Telco Customer Churn Prediction

Xinyuan Wang, Qiuyu Bao, Xiyi Fan

1 Introduction

In modern days, Telecommunications is undoubtedly one of the essential industries in our daily life, considering that we need to rely on their service to keep up with the fast-paced information and keep in touch with our beloved ones. The business of Telecommunication is undoubtedly very competitive around the world. For a company to stay in business and survive in the heated competition, one of the crucial things is to keep the customers. Thus, an important subject that needs to be done is predicting customer behavior, analyzing all relevant customer data, and developing focused customer retention programs. With the knowledge of statistics and data science, we aim to provide some analysis focusing on explaining what types of customers are more likely to stay and leave, creating prediction models to make customer churn predictions, and ultimately helping telecom companies survive and thrive among all competitions. This project report will provide an overview of our methodology, model selection, performance evaluation, and potential improvements for future iterations. We would first use exploratory data analysis to gain some intuitive information about the data set and the problem we are dealing with. Then we will mainly use two approaches to make user churn predictions. The first set of approaches involves using PCA (Principal Components Analysis) to reduce the dimension of the feature space and then using the k-means algorithm to cluster customers into different groups and estimate their churn-out rate based on their groups. The second set of approaches involves using a decision tree and random forest to make accurate forecasts and also seek the possible relationship of causation to user churn from the decision tree. We hope our report helps businesses enhance their decision-making process and maximize customer satisfaction.

2 Exploratory Data Analysis

The dataset we used for this project is sourced from IBM and contains comprehensive information on telecommunications customers, including their renewal status. The dataset consists of 7,043 entries with 21 columns, of which 19 are feature variables and only 11 instances have missing data. Our primary goal is to assist telecommunications companies in identifying the types of customers who are more likely to renew their service contracts. To achieve this, we initially explored the influence of various features on renewal decisions through exploratory data analysis.

To preprocess the dataset, we replaced the missing data with zeros. In this dataset, the proportion of customers who churned their contracts stands at 27.6%, while those who did not churn account for 72.4%.

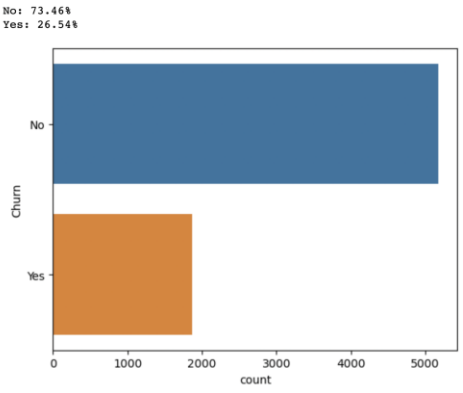

Figure 1: The proportion of renewal use

Next, we will classify the features into two categories: numerical data and categorical data and handle them separately.

2.1 Numerical Data

There are only three columns of numerical data in the table, namely subscription tenure, monthly charges, and total charges.

Considering that a histogram may not

provide a clear picture of the trend of the data, we will use Kernel Density Estimation (KDE) plots to show the trend. A higher probability indicates a more concentrated distribution of data. The KDE plots of subscription tenure, monthly charges, and total charges are shown below:

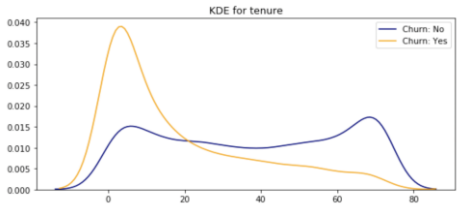

Figure 2: The KDE of Subscription Tenure

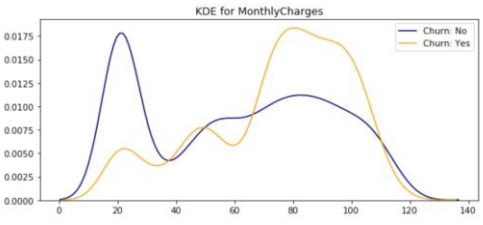

Figure 3: The KDE of Monthly Charges

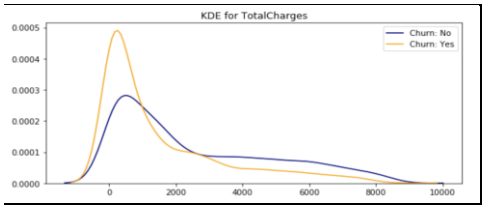

Figure 4: The KDE of Total Charges

From the Figure 2-4, we can analyze that:
(1) Users with shorter subscription tenure are more likely to churn their subscription.
(2) Users with higher monthly charges are more likely to churn from their subscription, while those with lower monthly charges are less likely to churn.
(3) Total charges do not have a significant impact on whether users will renew their subscription or not.
(4) Subscription tenure and monthly charges are two important features to consider.

2.2 Categorical Data

The dataset contains 16 categorical features, including 6 binary features, 9 ternary features, and 1 quaternary feature. For categorical features, we can use bar charts to analyze them.

(1) Binary Feature

We have selected four binary features, including age, partner status, family status, and paperless billing, to plot bar charts. The percentage of users will be plotted on the y-axis.

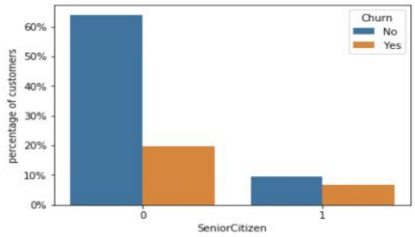

Figure 5: Relationship between age and renewal (0 means senior, 1 means not senior)

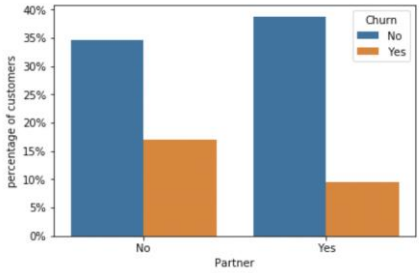

Figure 6: Relationship between whether the user has a partner and renewal

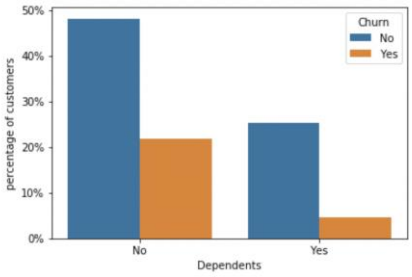

Figure 7: Relationship between whether the user has dependents and renewal

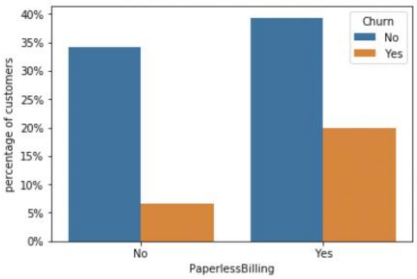

Figure 8: Relationship between paperless billing and renewal

Based on the Figure 5-8, we can draw the following conclusions:

1. Although the number of elderly residents is

small in the dataset, they are likely to churn their subscriptions.

2. Residents without family or partner are more likely to churn their subscription than those with family or partner.

3. Paperless billing is a payment method that is more likely to lead to subscription churn.

(2) Ternary Feature

To facilitate preliminary data exploration, we will separate the users into three categories: those who have both telephone and internet services, those who only have telephone services, and those who only have internet services. Figure 9 shows the impact of telephone services on subscription renewal:

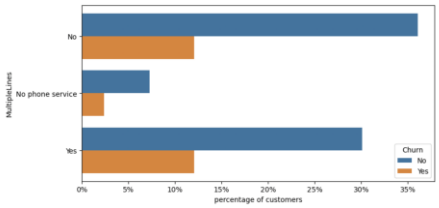

Figure 9: Relationship between phone service and renewal

It can be seen that users without telephone services are very few and users with multiple phone lines are more likely to churn their subscription than those with only one phone line, but the impact is not significant.

Figure 10 shows the impact of internet services on subscription renewal:

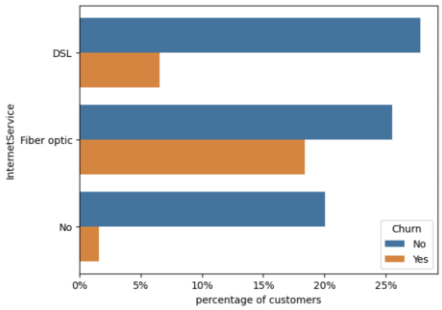

Figure 10: Relationship between internet service and renewal

From the Figure 10, we can see that unlike telephone services, about 20% of users do not have internet services. Users who subscribe to fiber optic services have a much higher churn rate compared to those who subscribe to DSL services. We can speculate that fiber optic services are worse than DSL services, which can make users to churn.

Internet services also have six additional services. Let's consider the impact of these six additional services on subscription renewal.

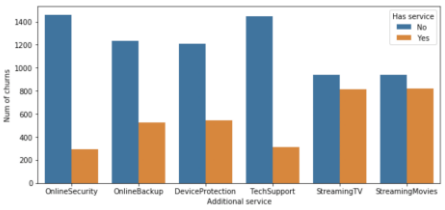

Figure 11: Relationship between additional service and renewal

From Figure 11, we can see that the first four additional services make users less likely to churn their subscription. The impact of the last two additional services on subscription renewal is quite negative.

The impact of contract length on subscription renewal is shown in Figure 12:

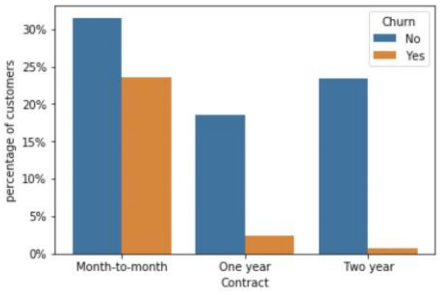

Figure 12: Relationship between contract length and renewal

From Figure 12, we can see that more than half of the users have chosen the month-to-month contract, and the churn rate is much higher than that of one-year or two-year contracts.

(3) Quaternary Feature

The dataset only has one quaternary categorical feature: payment method.

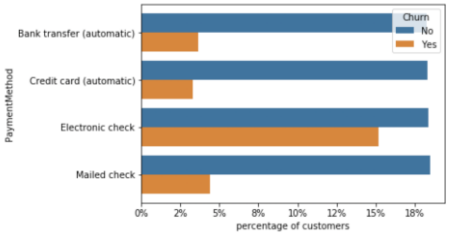

Figure 13: Relationship between payment method and renewal

Over 30% of users have chosen electronic check as their payment method, and the churn

rate is also much higher than the other three payment methods.

2.3 Feature Ordering

From the above analysis of the data, we can see that the impact of different features on subscription renewal varies. We can draw a heatmap to represent the relationship between the features.

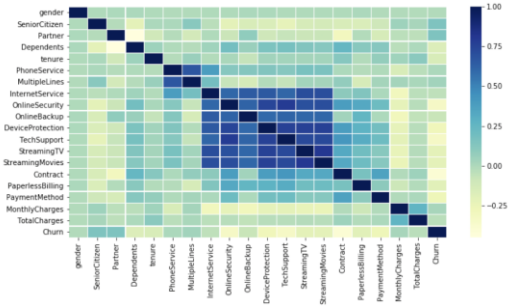

Figure 14: Correlation between variables

Excluding the diagonal relationship (because it is self-correlation and always has a correlation coefficient of 1), we can see from the heatmap that the highest correlation is between internet services and the six additional internet services. The correlation between other variables is relatively low, which is consistent with our intuition.

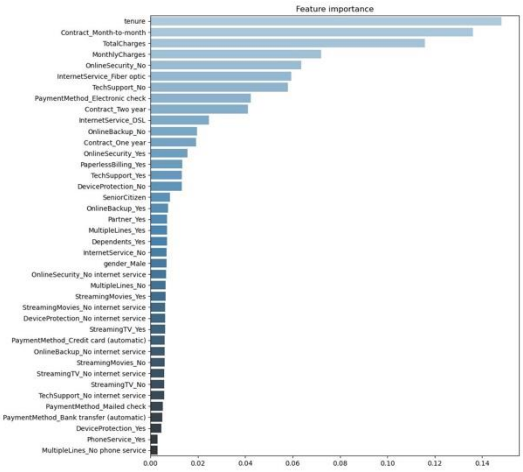

Figure 15: Feature importance ranking

From Figure 15, we can see that if we select subscription renewal as the research object, the most important features are contract tenure, total charges, monthly charges, type of internet services, and payment method. Preliminary data exploration shows that these features do have a significant impact on whether users will churn their subscription or not. Therefore, our main work in the next step is to explore the specific relationship between subscription churn and these features.

3 Dataset Dimensional Reduction

In the previous step, we sorted the feature importance of the data after filtering and cleaning regarding the churn target value. The importance of each feature shows a decreasing trend from high to low. The importance of the top three features, Tenure, Contract_Month-to-month, and TotalCharges, exceeds 0.1. The importance of the fourth-ranked feature, MonthlyCharges, is only 0.06. To avoid overfitting in K-means clustering and to deal with the potential correlation issues between some of the variables, we would attempt to use PCA algorithm to reduce the dimensionality to fewer dimensions for preprocessing. Before using PCA to reduce the dimensionality, we need to standardize the objective data except for customer_ID.

3.1 Data Processing

In our analysis, we organized the dataset columns into different groups, namely: ID, target, continuous features, and categorical features. Furthermore, we subdivided the categorical features into two types: binary values and those with more than two categories. To facilitate our modeling process, we implemented one-hot encoding for each categorical feature, creating separate sub-columns for individual categories and assigning a value of 1 if the category is present and 0 otherwise. Concurrently, we standardized the continuous features to ensure uniform scaling and reduce potential biases in the subsequent model training phase.

3.2 PCA

We computed the explained variances for

the top three principal components (PCs) using PCA, which were found to be 0.31, 0.20, and 0.07, respectively. Given that the third PC accounts for only 7.46% of the total variance, we opted to reduce the dimensionality to two principal components. It results in a dataset containing three variables and one target variable, with the two-dimensional PCA representation encapsulating the essential features of the original dataset.

Table 1: The dataset after PCA

	PC1	PC2	Churn
0	-1.31	-1.75	Not Churn
1	-0.32	-0.26	Not Churn
2	-1.17	-1.53	Churn
3	-0.12	0.35	Not Churn
4	-0.76	-2.39	Churn
…	…	…	…

3.3 K-Means

To evaluate the clustering efficiency, we employed the inertia metric, which measures the squared distance between centroids and data points. A lower inertia value indicates higher efficiency, as it suggests that the data points within each cluster are more concentrated and tightly grouped around their respective centroids. This metric helps us assess the quality of our clustering model and identify the optimal number of clusters for our analysis.

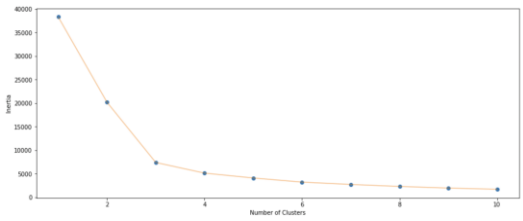

Figure 16: Inertia Decline Curve

According to the Figure 16, we choose the number of clusters to be 3 or 4.

Table 2: Calculate the clusters

	PC1	PC2	Churn	Clusters
0	-1.31	-1.75	Not Churn	2
1	-0.32	-0.26	Not Churn	0
2	-1.17	-1.53	Churn	2
3	-0.12	0.35	Not Churn	0
4	-0.76	-2.39	Churn	2
…	…	…	…	…

Following our analysis, we employed PCA and K-means clustering methods to effectively categorize customers into three distinct segments, each characterized by unique repurchase rates, which is estimated by averaging the churn out rate of all the customers within a cluster. When we are using 3 clusters to perform the k-means, we are getting the renewal rate of each cluster's customer to be 92%, 85% and 55%. Once we increase the number of clusters to 4, the renewal rate of each of the cluster is now 93%, 87%, 79% and 50%. Therefore, by classifying existing customer data, we were able to predict the probability of a customer making a repurchase, which in turn empowers businesses to make data-driven adjustments to their sales strategies.

4 Decision Trees and Random Forests

Upon establishing the PCA and k-means models for customer segmentation and the development of improved communication strategies, we furthered our classification efforts using decision trees and random forests. This should help us not only establishing a model with reasonably good accuracy, but at the same time, it would also provide us some insights on what are the features that influence the user churn out rate the most. We created decision trees based on three distinct feature sets: 1) all features, 2) subscription status for phone and internet services, along with contract duration, and 3) senior citizen status, subscription tenure, and monthly charges. The second feature set can be viewed as a set that contains the characteristics of the service. The third feature set contain features that are regarding the characteristics of the customers. We then constructed ROC curves for each tree at specific AUC values to evaluate the performance of the decision tree and random forest models. This additional layer of analysis

allows us to refine our customer classifications and develop more targeted and effective communication strategies.

4.1 Data Processing

For all categorical variables in the dataset, we applied a numerical encoding process to transform them into distinct integer values. For binary variables, we assigned values of 0 and 1. For variables with three categories, we assigned values of 0, 1, and 2. Finally, for variables with four categories, we assigned values of 0, 1, 2, and 3. This encoding process simplifies the handling of categorical variables and facilitates their integration into our machine learning models.

4.2 ROC Curve of All features Tree

Figure 17 provides the information of the decision tree that includes all features. This decision tree has an accuracy of 72%. An ROC curve, or receiver operating characteristic curve, is a graphical representation that illustrates the effectiveness of a classification model across all possible classification thresholds. This graph plots two crucial metrics: the True Positive Rate and the False Positive Rate. By reducing the classification threshold, more instances are identified as positive, which subsequently leads to a rise in both False Positive and True Positive rates. The Area Under the Curve (AUC) signifies the likelihood that a randomly selected positive instance (labelled green) is positioned to the right of a randomly selected negative instance (labelled red). The AUC value fluctuates between 0 and 1. A model with entirely inaccurate predictions has an AUC of 0.0, while a model with entirely accurate predictions has an AUC of 1.0. AUC is scale independent. It evaluates the ranking of predictions, rather than their absolute numerical values. AUC is also invariant to the classification threshold. It assesses the model's predictive quality regardless of the selected classification threshold. The AUC of our decision tree is 0.64, which is not bad.

Based on the decision tree we created, among all variables, the purchase of phone services has a direct impact on subsequent phone-related services, while the purchase of internet services directly affects subsequent internet-related services. Customer-specific attributes, such as senior citizen status, also influence the adoption and usage of these services, which in turn determines monthly charges. Therefore, we approached the analysis from the customer's perspective, examining how the subscription of phone and internet services, along with contract duration, influences the decision to renew.

```
The accuracy of the Decision Tree is 0.7232227488151659
```

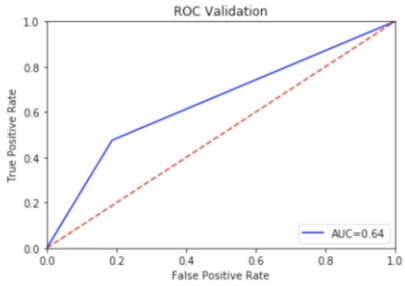

Figure 17: ROC Curve

4.3 Analysis of Decision Trees

As observed in Figure 18, the decision boundaries obtained under this classification scheme are not optimal, which may be attributed to the limited impact of monthly or annual contracts on the decision to renew. Consequently, we adopted the third approach for constructing decision trees and decision boundaries, focusing on the customer's inherent characteristics. We assessed whether the customer is a senior citizen, their subscription tenure, and monthly expenses to determine the likelihood of renewal.

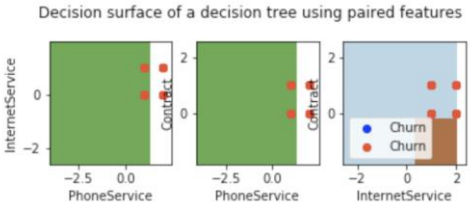

Figure 18: Decision surfaces plotted by phone, internet service

and contract length

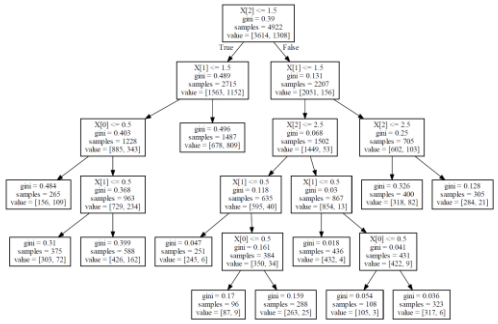

Figure 19: Decision tree plotted by phone, internet service and contract length.

Based on Figure 20, we can observe that senior citizens are less likely to renew their contracts if they have already been subscribed for an extended period, while younger individuals tend to be more willing to renew after trying the service for a while. For customers with longer subscription tenures, a higher monthly charge is associated with a smaller likelihood of renewal. This could potentially be interpreted as these customers having developed a dependence on the communication services provided, that even the charge is high, they would still like to stay.

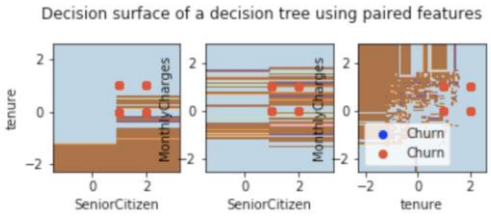

Figure 20: Decision surfaces plotted by senior, monthly charges, and tenure.

4.4 Data Remedy and New Tree Analysis

In the previous Decision tree process, we noticed that the accuracy rate according to the Decision tree model is always around 70%. This finding alerts us since we noticed that this is the same as our user renewal rate. As we see in Figure 19, all the leave nodes of the tree are predicting that the user is going to stay except for one. This problem is likely to associate with our imbalanced data set. To fix this problem, we randomly select the same number of customers that churned and renewed, so that our data set is now more balanced target wise.

After our balanced our data, the all-feature decision tree's accuracy dropped to 69%, yet the AUC has increased to 0.69, which indicate that our model now does not suffer much from the unbalanced data set.

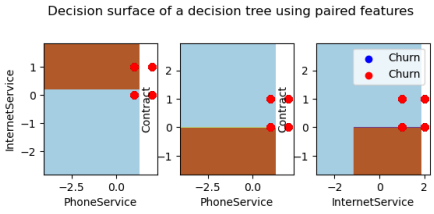

Figure 21: Decision surfaces plotted by phone, internet service and contract length with better data.

Figure 21 is the same decision surface as Figure 19 with the balanced data set. We can now see that longer contract would almost guarantee the renewal of service. This tree has an accuracy of 74%.

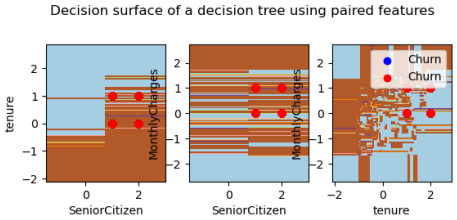

Figure 22: Decision surfaces plotted by senior, monthly charges, and tenure with better data.

Figure 22 is the same decision surface as Figure 20 with the balanced data set. We can now see that now there are stronger evidence showing that high monthly charges make people across all age group less willingly to renew, unless they have already been with the company for a long time. This tree has an accuracy of 68%.

Figure 23: Decision surfaces plotted by senior, Partner, and

Dependents with better data.

We would also like to make some analysis on the decision surface of the customer's statistics only, so what plot the following decision surface in Figure 23. As we can see, senior citizens are likely to churn in general, except for when they have dependents. Younger age group are likely to churn if they have no partners and dependents. Our results demonstrated that people will likely renew when they have more things and people that they are caring.

4.4 Random Forest

Random Forest is a machine learning algorithm that is preferred over a single Decision Tree because it reduces overfitting, produces more accurate predictions, is less sensitive to outliers, can handle missing values, and provides a measure of feature importance. Overall, it is a powerful algorithm that can provide better performance, especially when dealing with complex or high-dimensional data. To make our prediction more accurate and to reduce the variance of our decision tree model, we would like to implement the random forest method.

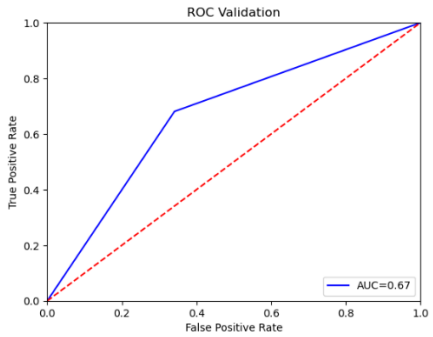

Figure 24: ROC for Random Forest

We see in Figure 24, the AUC is still high, and the accuracy rate is now around 75%.

5 Summary

Our current analysis indicates that decision trees and random forests utilizing a voting mechanism may provide a superior performance compared to K-means clustering especially in terms of interpretability. While the K-means method offers a straightforward implementation, it may lack the necessary precision for accurate predictions. If the differences in repurchase rates between the segments are minimal, the prediction quality may be negatively impacted, ultimately resulting in less effective outcomes.

We could develop more products targeted specifically at the group likely to renew such as varies family plans to help people connect with their families, which has been shown as a great need in our report. Tailoring products to this group's preferences and needs could drive customer satisfaction and loyalty.

Continued feedback collection from customers is crucial. With advancements in Natural Language Processing (NLP), we might be able to extract and analyze customer sentiments more effectively, providing deeper insights into their experiences and wish.

Exploring better datasets can also prove beneficial. More comprehensive and diverse data can improve the performance of our prediction models and provide more nuanced insights into customer behavior. We have managed to achieve this by balancing the data set, but more data might be needed.

It would also be very helpful if we can gain more features from the data set. Our data focuses more on the family structural information of the customers instead of their occupation and income. We believe obtaining those data can help us target better those business or professionals who requires good quality communication and willing to pay more to the services, which ultimately could help the profit of the company.

By addressing these directions, we can continue to refine our churn prediction models, enhance customer engagement strategies, and ultimately drive the success of tele-communications companies.