CS610 Applied Machine Learning

Final Report

A Prediction in Telecom Customer Churn

Group 4

Chen Fangxian

Chen Yiman

Tan Hui Ying, Amanda

Tay Zhi Xiang

Quek Zixin, Richard

1. Project Objective

Churn is a widespread problem across almost every industry, acquiring new customers can cost 5-7 times more than retaining existing customers (Kumar, 2022). Traditionally, company turns to agency for survey results to understand customer behaviours, however, with vast amount of data that a company has from customers, we could make use of machine learning to analyse features that might cause churn from within. This could help businesses to understand their customer churn to better strategize and implement campaigns that are targeted to customers.

This project aims to help to telecom companies to understand which factors are important contributors that cause customers' churn. We ranked the factors from most important to least important so companies will be able to deep dive into individual component of the business to address the churn.

In this project, 4 machine learning models are selected and trained to carry out the binary prediction of if a customer churn or not churn: Logistic Regression, Random Forest, Gradient Boosting, Multi-Layer Perceptron. We then pick up model with best performance to conduct hyperparameter tuning for further improvement.

2. Dataset and Pre-processing

a. Dataset Overview

The dataset was retrieved from Kaggle, a sample data from International Business Machines Corporation (IBM), which offers a diverse sample of telecommunication customers across various cities in California over a one-month period. Encompassing 7,043 observations (rows) and 33 features (columns), this dataset would allow us to gain comprehensive insights into customer behaviour and preferences.

To preserve the unique characteristics of each city, the median income per capita figures was retrieved from the US Census, and we matched to the cities by using Excel VLOOKUP.

As stated in the project objective, the target variable is "***Churn Value***", while the independent variables include the remaining variables from the dataset, however, we carried out some data exploratory analysis and pre-processing to exclude the irrelevant ones.

b. Data Exploratory and Pre-processing

To enhance the quality, consistency, and usability for our predictive modelling, it is vital to pre-process the dataset prior to engaging in the customer churn analysis. Below are the steps taken:

i. **Irrelevant Data**

Figure 1 shows the 20 Churn Reasons and service standards and competitors stand out as stronger reasons for customers to churn. However, there are some reasons such as "Moved" and "Deceased" are inevitable for customers to stop the services. Hence, we removed these corresponding rows from the dataset.

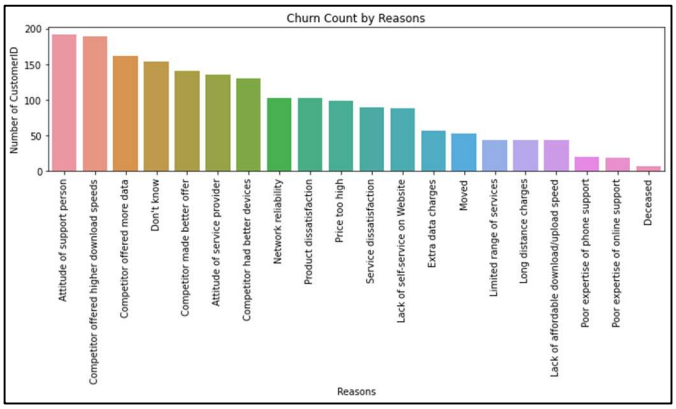

Figure 1

ii. **Imputation**

There are some missing data for "Total Charges" and "income per Capita", median data is imputed to these two variables to maintain its overall distribution.

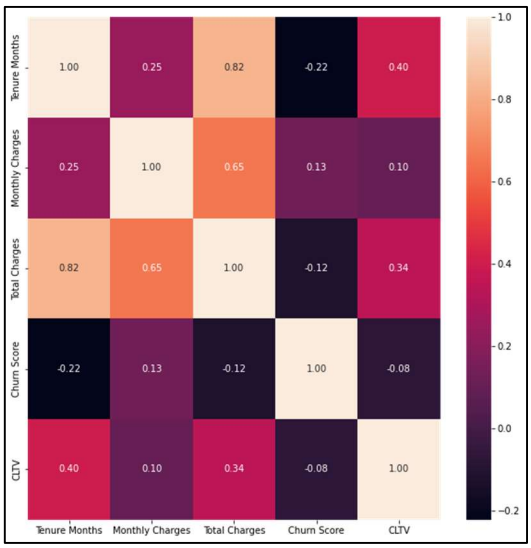

Figure 2

iii. **Multicollinearity**

To understand more on how continuous variables are related within this dataset, multicollinearity heatmap was plotted for the analysis as shown in Figure 2.

Figure 2 shows that "Total Charges" are highly correlated with "Tenure Months" and "Monthly Charges". We noted on this and removed "Total Charges" run logistic regression separately in the later steps to find out if it helps to improve the performance. For the other models, we include "Total Charges" as Multicollinearity does not affect its performance.

iv. **Drop irrelevant columns**

All columns are investigated to think about how it could help for the churn prediction, the following columns are dropped as they will not contribute for the prediction: CustomerID, Count, Country, State, City, Zip code, Lat Long, Latitude, Longitude, Churn Score, Churn Reason.

v. **One-hot encoding**

There are many columns are categorical data such as Gender, Dependent, Payment Method, etc. these are encoded to numerical data to range from 0-4 for model running later.

vi. **Standardization**

After encoding, the continuous variables are completely out of scales. This will cause inaccuracy in running the models, standardization was then carried out to make these columns to the range of the similar scale.

vii. **Imbalanced data**

The proportion of churn label "No" is three times more than "Yes", this will impact the prediction, especially in this project where binary prediction is the objective. SMOTE (synthetic Minority Oversampling Technique) was used before we split the data into training and test sets to generate new instances from existing data in a balance way.

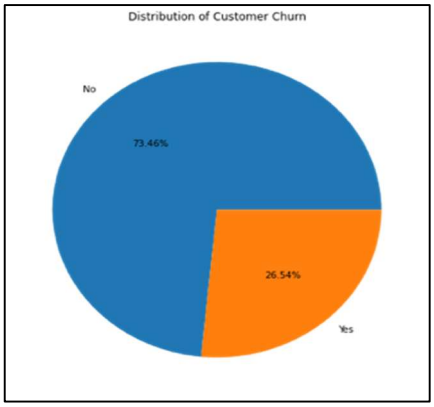

Figure 1

c. **Final Used Features**

Below table showing the final features we have used for the models

Type	Features	Remarks
Discrete	Gender, senior citizen, partner, dependents, tenure months, phone service, multiple lines, internet service, online security, online backup, device protection, tech support, streaming movies, contract, paperless billing, payment methods	One-hot encoded based on the original values
Continuous	Tenure, Monthly charges, Total Charges, CLTV, Income per capita	The continuous features were standardized

3. Model Building

a. **Logistic Regression (LR)**
LR is a supervised machine learning algorithm that accomplishes binary classification tasks by predicting the probability of an outcome, event, or observation. LR analyzes the relationship between one or more independent variables and classifies data into discrete classes. One of the pros of LR is it could provide valuable insights: LR measures how relevant or appropriate an independent/predictor variable is (coefficient size) and reveals the direction of their relationship or association (positive or negative). For this project, we applied LR to predict the probability of churning and then assigned 1/0 label for classification.

b. **Random Forest (RF)**
RF model is made up of multiple decision trees, which start with a basic task .These questions make up the decision nodes in the tree, acting as a means to split the data. Decision trees seek to find the best split to subset the data, and they are typically trained through the Classification and Regression Tree (CART) algorithm. To overcome problems such as bias and overfitting, random forest was ensembled for more accurate results. In our project, RF reduced risk of overfitting, bias and overall variance thus gave us a more precise prediction for churning.

c. **Gradient Boosting (GB)**
Boosting is a powerful ensemble technique in machine learning, which combines the predictions of multiple weak learners to create a single, more accurate strong learner. **One of the advantages of GB** is the flexibility: it can be used in many important tasks such as regression, and classification. Besides, another pro of GB is the interpretable: unlike black-box algorithms like neural networks, GB does not sacrifice interpretability for performance. It works like a Swiss watch and yet, with patience, you can learn and teach others how it works to others easily. In our project, GB formed a strong learner and reduced risk of overfitting, bias and overall variance thus gave us a more precise prediction for churning.

d. **Multi-layer Perceptron (MLP)**
The MLP is a neural network where the mapping between inputs and output is non-linear. A MLP has input and output layers, and one or more hidden layers with many neurons stacked together. Each layer is feeding the next one with the result of their computation, their internal representation of the data. This goes all the way through the hidden layers to the output layer. For our prediction task, we built hidden layer with ReLU as activation function to predict the chance of discrete classification.

4. Results and Analysis

Comparing machine learning (ML) models is crucial for predicting customer churn while ensuring it can generalize well with new and unseen data. After splitting the dataset into training (80%) and testing (20%) with *np.random.seed (2004)* and applying across all models, the results are analysed and plotted against both Area Under Precision-Recall Curve (AUPRC) and Receiver Operating Characteristics (AUROC) for each ML models considered.

a. **Model Comparison**

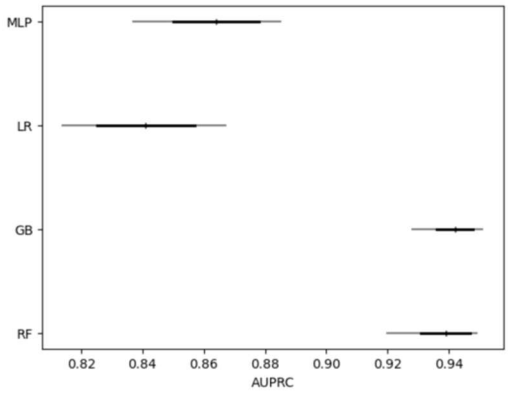

Figure 2

	Model	Threshold	AUROC	AUPRC	Sensitivity	Specificity	Runtime
0	RF	0.580	0.942 (0.928-0.949)	0.939 (0.92-0.949)	0.851 (0.846-0.911)	0.885 (0.823-0.897)	4.934
1	GB	0.543	0.936 (0.925-0.943)	0.942 (0.928-0.951)	0.85 (0.835-0.874)	0.862 (0.828-0.882)	2.219
2	LR	0.539	0.864 (0.85-0.877)	0.841 (0.814-0.867)	0.802 (0.77-0.831)	0.78 (0.755-0.826)	0.299
3	MLP	0.455	0.89 (0.875-0.899)	0.864 (0.837-0.885)	0.86 (0.819-0.869)	0.783 (0.765-0.811)	6.169

Figure 3

The following interpretation of the results is presented with rank "1" as the best outcome in their respective category:

- **Threshold** – ability to represent the bias in making prediction.
- **AUROC** – ability to differentiate between positive and negative instances.
- **AUPRC** – ability to identify positive instances while minimizing false positives.
- **Sensitivity (Recall)** – ability to correctly classify actual positive instances.
- **Specificity** – ability to correctly classify actual negative instances.

	Threshold	AUROC	AUPRC	Sensitivity	Specificity	Score	Rank
RF	1	1	2	2	1	**7**	**1**
GB	2	2	1	3	2	10	2
LR	3	4	4	4	4	19	4
MLP	4	3	3	1	3	14	3

The area under the AUPRC is a superior metric for model comparison to the area under the AUROC for binary classification tasks with class imbalance. As we have pre-processed with the imbalanced data, we chose AUROC rather than AUPRC as performance measures. For AUROC, RF take the first place. So we chose RF to tune hyperparameter to get higher accuracy. However, after tuning *max_depth, n_estimators,* and *random_state*, the AUPRC decreased to 0.91.

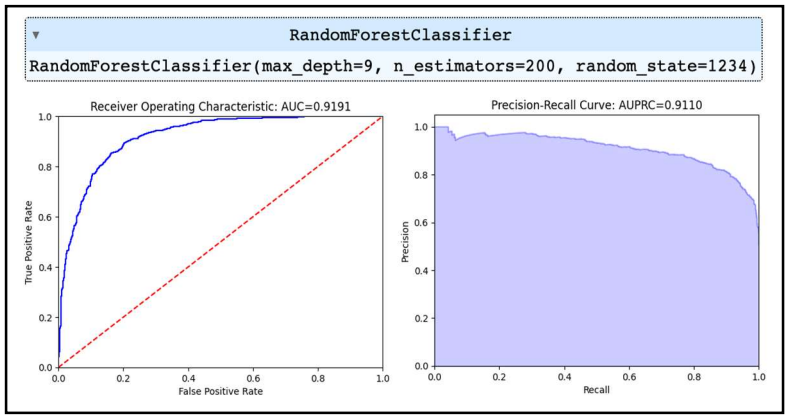

Figure 4: RF-ROC-AUC: 0.942 -> 0.9191 (Decreased)

GB model took the second place, we also tried to tune hyperparameter to see if it can reach a higher AUROC. We *tuned max_depth, max_feature, min_samples_leaf, min_sample_split, n_estimator* and *random_state*, finally the tuned GD model performance was increased to 0.9509, which is the highest up to now.

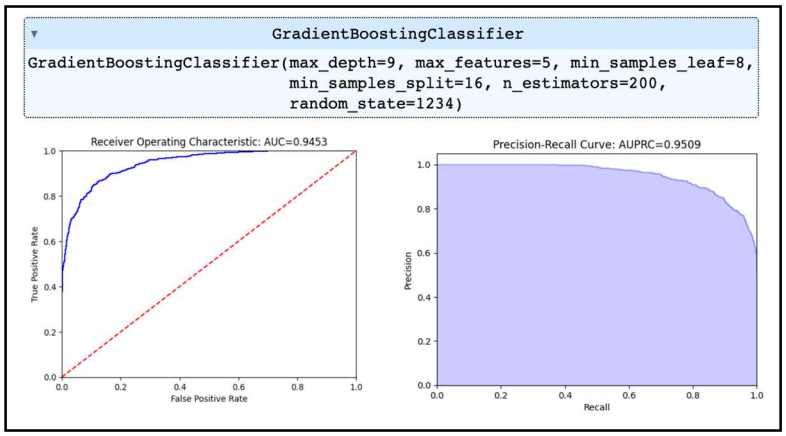

Figure 5: GB-ROC-AUC: 0.936->0.9453 (Increased)

b. **Feature Importance**

Feature importance is a step in building a machine learning model that involves calculating the score for all input features in a model to establish the importance of each feature in the decision-making process. The higher the score for a feature, the larger effect it has on the model to predict a certain variable.

Hyperparameter tuning has not only optimized the performance of our customer churn prediction models but also highlighted on the most influential features that determine attrition. Feature importance analysis from the most accurate model – Tuned Gradient Boost is analyzed to identify the most important feature affecting churning or not. The most important features are recorded in Figure 8:

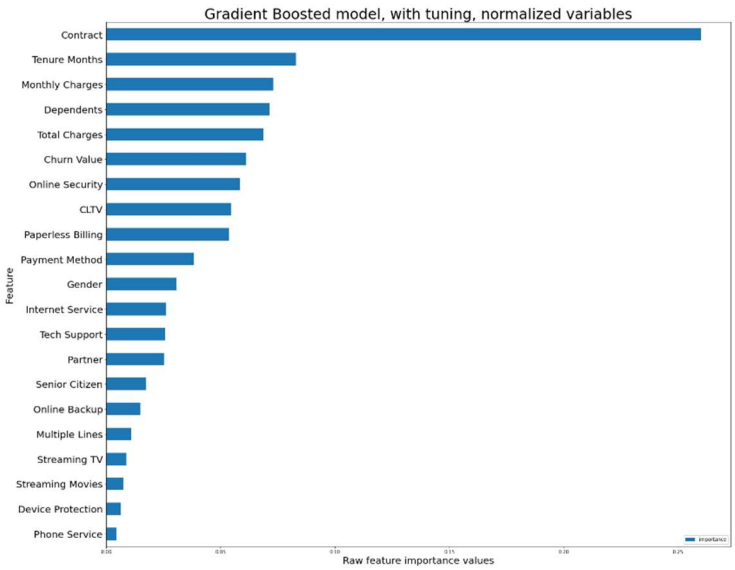

Figure 6

From figure 8 highlights that contract type of month-to-month subscribers are more likely to churn due to their reduced contractual obligations if they decide to terminate their service. This insight would guide strategies to improve customer retention for this group.

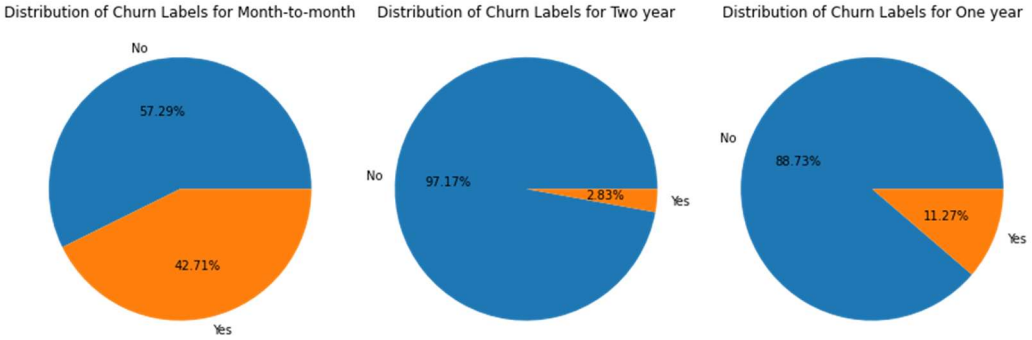

Figure 7: distribution of Contract Type by Churn Label

5. Conclusion, Limitation and Future Works

Our analysis has identified crucial predictors of customer churn, using a RF model that initially attained the lowest False Positive Rate (FPR) and predicted the highest ROC. Following tuning, we achieved even higher ROC with a tuned GB model.

Four primary features emerged as significant churn indicators: **Contract**, **Tenure Months**, **Monthly Charge** and **Dependent**. These findings enable us to develop targeted strategies aimed at enhancing customer retention and satisfaction. Organization can effectively minimize customer churn, promote loyalty, and drive organizational growth and profitability by acting on these predictive features.

While our analysis has provided valuable insights into customer churn prediction, further enhancement to model performance can be done by addressing the following limitations:

i. **Limited dataset duration** – The current dataset spanning only one month may have restricted the robustness of the model. Access to an extended timeframe would enable the model to capture more complex patterns and better adapt to changing trends.

ii. **Absence of timestamp data** – The lack of timestamp information may have limited the model to make more accurate churn predictions based on timely factors. Integrate this datapoint would allow the model to identify key churn drivers and address issues proactively and timely.

Building upon our analysis and its key findings, several avenues for future work have been identified that could significantly enhance our understanding of customer churn and retention:

i. Perform an in-depth exploration and analysis on the most critical features highlighted through feature importance to better understand the correlation between these factors and customer churn. This targeted approach would strategically allocate resources more effectively to retain customers and improve overall customer satisfaction.

ii. Integrate additional contextual factors such as boarder telecommunications competitor data which are accessible through data agencies with a cost. This would provide greater insights to churn dynamics and trends to make informed strategic decisions.

iii. Consider inclusion of promotional mechanisms and their impact on customer churn. More effective strategies can be crafted to reduce churn, optimize retention efforts, and drive long-term loyalty through examining how promotions influence customer behavior.

6. Reference

Kumar, S. (2022, December 12). *Customer Retention Versus Customer Acquisition*. Forbes. Retrieved March 10, 2024, from: https://www.forbes.com/sites/forbesbusinesscouncil/2022/12/12/customer-retention-versus-customer-acquisition/?sh=28ce8a6f1c7d