Subscription Service Customer Churn Prediction

ECS 171 Machine Learning
Group 3 Project Report

Group members:

Sajid Hussein, Bradley Phyo, Muhammad Reza, Zeshang Wang

Github repository:

https://github.com/Hackingsimulator/ECS_170

Introduction and Background

The subscription service industry is highly competitive, with customer churn being a major concern for companies. ***Churn***, the rate at which customers discontinue their subscriptions, can have a significant impact on a company's revenue and growth. Predicting and understanding churn behavior is crucial for businesses to retain their customers and improve their service offerings.

Our project aims to address this challenge by developing a machine learning model that predicts churn rates within a specific subscription service.We will analyze a range of parameters, including the type of services customers are subscribed to, their payment methods, whether they have a partner or not, and their overall usage habits and the amount of money they spend. By examining these variables, we can uncover patterns and insights that contribute to customer churn.

The dataset we will utilize is from American multinational technology corporation IBM which contains individual customer observations, including their subscription details, payment history, and behavioral patterns. By leveraging this data, we can train our machine learning model to accurately predict the likelihood of a customer churning.

The primary objective of our project is to build a robust churn prediction model that can identify customers at risk of discontinuing their subscriptions. By doing so, we aim to empower businesses in the subscription service industry with the ability to proactively address customer churn and take appropriate actions to retain their valuable clientele.

Once our model is developed and validated, it holds the potential to provide valuable insights and benefits for various stakeholders within the subscription service industry:

● Customer Retention and Engagement:
- Companies can utilize the churn prediction model to identify customers at a high risk of churning.
- This information enables targeted retention efforts, such as personalized offers, tailored communication, or improved customer support, to enhance customer satisfaction and prevent churn.

● Service Optimization:
- Insights gained from the model can aid in optimizing service offerings. By understanding the specific services that correlate with higher churn rates, companies can improve or modify those offerings to better align with customer preferences and expectations.

● Marketing and Campaign Management:
- The churn prediction model can guide marketing strategies and campaign management. Companies can allocate resources more effectively by focusing on customers identified as at-risk of churning, ensuring marketing efforts are targeted and impactful.

Literature Review

Machine learning techniques are increasingly utilized in customer retention strategies to predict customer churn, which is crucial for organizations in highly competitive service sectors. Sahar F. Sabbeh's study compared different machine learning techniques for customer churn prediction using data from a telecommunications company. The study found that Random Forest and Adaptive Boosting performed the best, accurately predicting customer behavior about 96% of the

time. Other techniques such as Multi-layer perceptron and Support Vector Machines also achieved high accuracy rates. Similarly, Praveen Asthana conducted a comprehensive study comparing machine learning methods for customer churn prediction. The study revealed that BPN and DT were the top-performing classifiers, while SVM with Radial Basis Function and Polynomial kernels also performed well. Boosting techniques significantly improved the accuracy of certain classifiers, and the SVM-POLY using AdaBoost achieved an accuracy of nearly 97%. These studies highlight the effectiveness of machine learning in predicting customer churn and offer insights for shaping customer retention strategies.

In "How Machine Learning Can Help with Customer Retention" by Euge Inzaugarat, the importance of customer churn prediction models for e-commerce and online businesses is emphasized. The paper discusses the significance of churn prediction models in identifying significant customer features related to churn. Using a dataset of 10,000 bank customers, the author observed that factors such as gender, nationality, membership status, and the credit score by age ratio played a significant role in customer churn. The paper evaluated three models: Logistic Regression, Support Vector Machine (SVM), and Random Forest, with Random Forest performing the best with an accuracy of 84%. Feature importance analysis revealed that gender, number of products, member status, and the credit score by age ratio were the most influential variables. This research provides valuable insights into applying machine learning for customer churn prediction and informing customer retention strategies.

Dataset Description and Data Analysis

The dataset used for this project is the TelcoCustomerChurn dataset in the form of a csv file. Before data cleaning, there were approximately 7043 rows and 21 columns in the dataset. If the “Churn” column is disregarded since that is the dependent variable, there are 20 columns that describe a specific customer or subscriber. 6 out of 20 of these columns depict the different attributes of a customer (gender, customer ID, Senior Citizen, Partner, Dependents, tenure), 5 out of 20 of these columns depict their phone and internet habits (Phone Service, MultipleLines, Internet Service, Online Security, Online Backup), 4 out of 20 of these columns depict the types of services they are part of (Device Protection, Tech Support, Streaming TV, Streaming Movies) and lastly 5 out of these 20 columns depict their monetary habits (Contract, Paperless Billing, Payment Method, Monthly Charges, Total Charges).

After doing some exploratory data analysis, 5 variables or columns were dropped due to insufficient or minimal correlation to the outcome of a customer Churning. These 5 variables were Customer ID, Tenure, Partner, Gender, and PaperlessBilling. Out of the 15 variables left, 13 of them were categorical binary variables with only “Yes” or “No” as the possible attributes while Total Charges and Monthly Charges were the only variables that were numerical. The categorical variables were then changed to integer values via one-hot encoding.

Overall, this was a fairly straightforward dataset and contains easy to understand information and variables. There was not much to process except for taking out some NaN values and dropping unrelated variables.

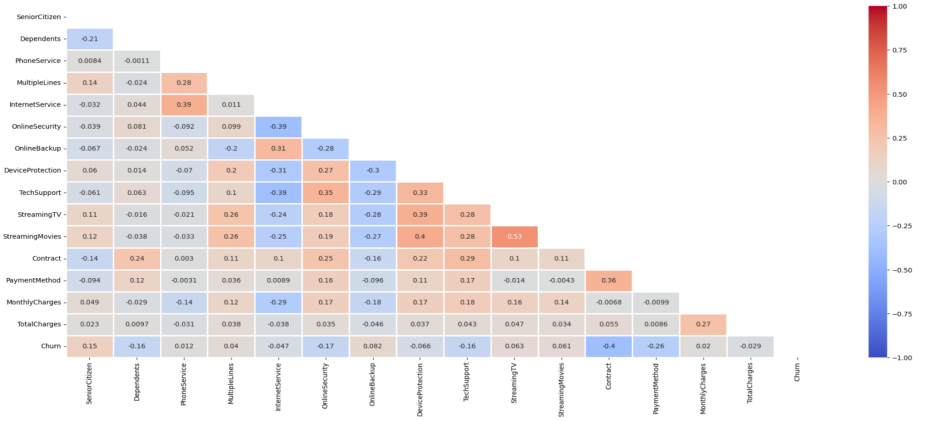


Figure 1. Heat map of dataset

Proposed Methodology

The problem at hand is a binary classification problem. The main challenge we faced was that the dataset contained an immense number of attributes that could influence whether a customer stays or leaves; hence, a more complex prediction model was required. From the papers we studied, we concluded that Feed-Forward Neural Networks (FFNN), Random Forest (RF) decision trees, Support Vector Machine (SVM), and Logistic Regression (LR) are the most popular models for customer retention-related problems.

The goal is to create a user interface that is accessible to any company. This interface will help determine customer retention factors, with the aim of optimizing profits and ensuring long-term success. The user interface accepts customer information as input and produces a prediction based on the model trained in the backend. The creation of the interface was straightforward and it works efficiently.

In addition, the results were looked over to understand which features mattered the most. The goal is to figure out which feature a streaming company should focus on and prioritize to keep their customers and maintain their profits. Ultimately, this approach would significantly aid in effective customer retention for the streaming service.

FNN:

	precision	recall	f1-score	support
0	0.73	1.00	0.85	1033
1	1.00	0.00	0.00	374
accuracy			0.73	1407
macro avg	0.87	0.50	0.42	1407
weighted avg	0.80	0.73	0.62	1407

Figure 2. Classification report of FNN

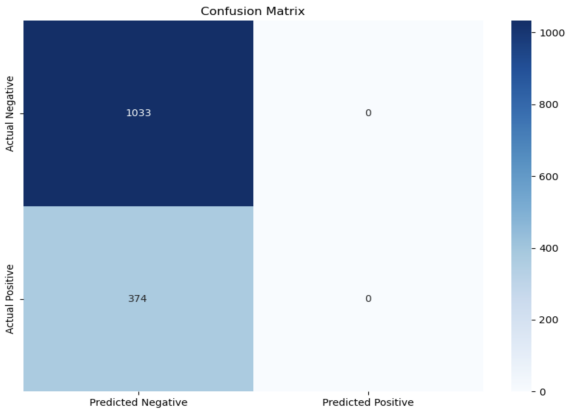


Figure 3. Confusion matrix of FNN

The first model that was implemented was MLPC, which is a specific type of FNN. Grid search was also utilized to determine the best hyperparameters to use for hyperparameter optimization. The libraries utilized were GridSearchCV from scikit-learn and MLPC Classifier from sklearn.neural_network.

There were two hidden layers sized 10 and 13 respectively. The learning rate was 0.3 and the maximum interactions were 500. Overall, the optimal accuracy was 73%. After running the training and testing data multiple times, it was found that the FNN model too had an overall accuracy of ***73%.***

While this model was straightforward, it was not optimal as the confusion matrix was not as clear as the other models and often led to errors in the code.

Logistic Regression:

	precision	recall	f1-score	support
0	0.83	0.89	0.86	1033
1	0.61	0.50	0.55	374
accuracy			0.78	1407
macro avg	0.72	0.69	0.70	1407
weighted avg	0.77	0.78	0.77	1407

Figure 4. Classification report for Logistic Regression

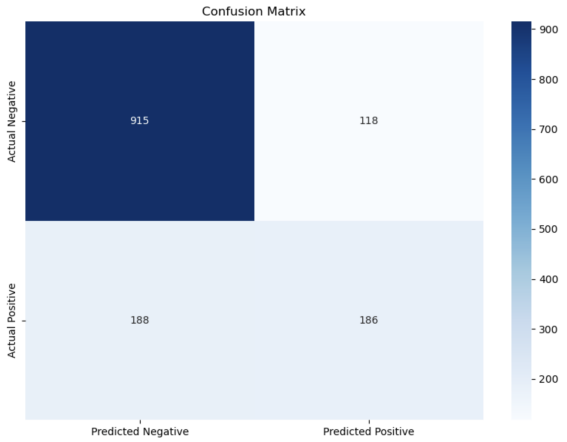


Figure 5. Confusion Matrix for Logistic Regression

The logistic regression model was built using the LogisticRegression function from the sklearn library in Python, and set with a maximum iteration parameter of 10000 to ensure convergence. The model was trained using the training dataset and tested. By using sklearn's metrics.accuracy_score function the accuracy is measured to be ***78%***. Therefore, the model provides a relatively accurate prediction of whether a customer is going to stay or leave which is the best model in all of the models used. The amount of predicted positive to actual positive was 186 while the amount of predicted positive to actual negative was 118. The amount of predicted negative to actual positive was 188 while the amount of predicted negative to actual negative was 915.

Decision Tree/Random Forest Classification:

	precision	recall	f1-score	support
0	0.81	0.80	0.81	1033
1	0.47	0.49	0.48	374
accuracy			0.72	1407
macro avg	0.64	0.64	0.64	1407
weighted avg	0.72	0.72	0.72	1407

Figure 6. Classification report for DT/RF

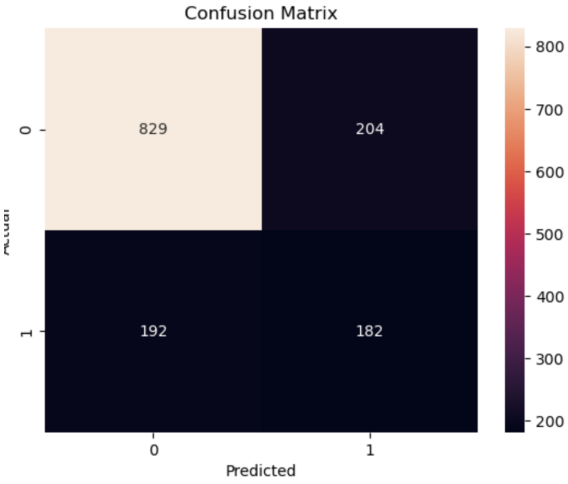
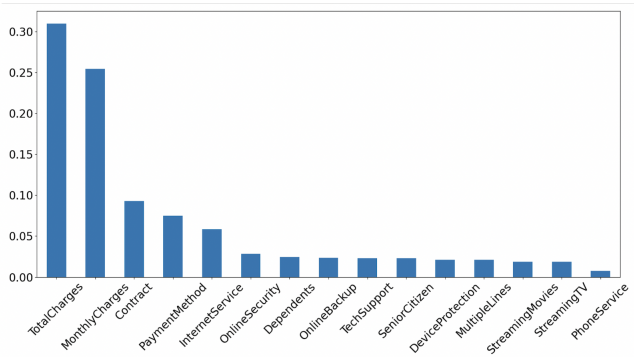


Figure 7. Confusion matrix for DT/RF

For decision trees and random forest classification, it was first thought that it would have the highest accuracy rate since most of the data was binary and decision trees are often the best model for binary data. However, it only had an accuracy of ***72%*** after running the testing and training data. From Random Forest Classification the model was also used to figure out the most important attributes and variables in this model which were the amount of money a customer spent that correlates to them churning.



Support Vector Machine (SVM):

	precision	recall	f1-score	support
0	0.73	1.00	0.85	1033
1	1.00	0.00	0.00	374
accuracy			0.73	1407
macro avg	0.87	0.50	0.42	1407
weighted avg	0.80	0.73	0.62	1407

Figure 9. Classification report for SVM

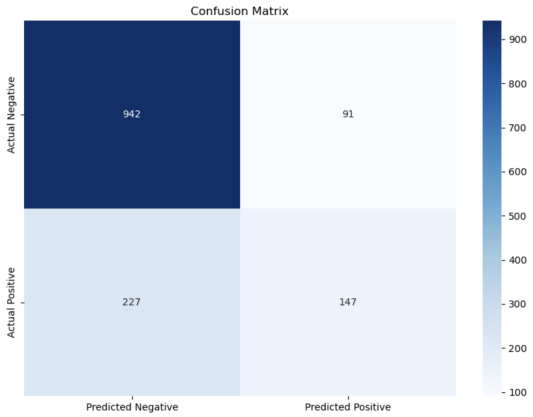
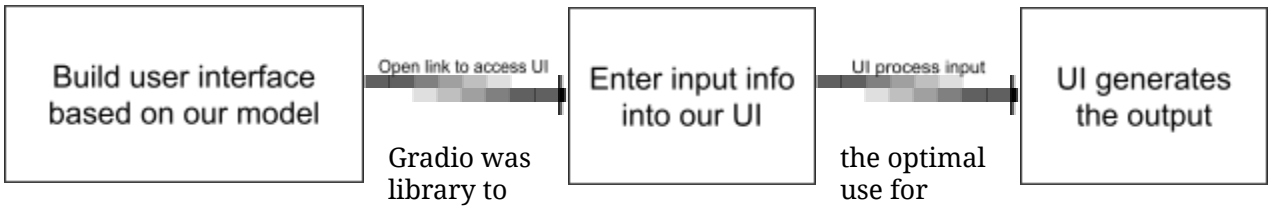


Figure 10. Confusion matrix for SVM

The SVM model was implemented with a linear kernel. The reason a linear kernel was used was because the variables were binary and very straightforward. It was more computationally efficient and worked better because the classes were easy to separate. Although it yielded a relatively high accuracy of ***73%,*** the logistic regression model was slightly better. By using SVM, the model was able to determine the best hyperplane that can optimize margin between the classes. The amount of predicted positive to actual positive was 147, while predicted positive to actual negative was 91. The amount of predicted negative to actual positive was 227 while the amount of predicted negative to actual negative was 942.

Building the User Interface

After determining the best model to use for predicting Churn, the next step was to build a user interface based on our best model for demonstration purposes. There are lots of very convenient and standard libraries in Python for building user-friendly UIs. The most ideal one for any project would be one that meshes well with our data, which are mostly categorical. Through extensive research on many different libraries, Gradio was determined to be the most optimal library to use for the UI. Below this paragraph is the general framework for how the UI operates.



Gradio was library to multiple

the optimal use for reasons.

The first reason is that the display of the UI that Gradio generates is very simple and easy to navigate, making the demonstration accessible for any and all users.

The second and most important reason that Gradio was so convenient is because of the flexibility it offers when dealing with categorical data. There are many convenient input functions, such as Gradio.Dropdown(), Gradio.Radio(), and Gradio.CheckboxGroup(), that are absolutely perfect for building a UI that processes categorical data. There are also keywords for input, such as “checkbox” that allow users to simply check off a box as they are entering the input to represent “yes” or 1, or “no” or 0. This keyword was especially useful for the purpose of our data, since a lot

of the categorical attributes in the dataset were binary. For the categorical attributes that had more than two categories, the Gradio.Radio() function was utilized, which allows users to choose one out of at least three options. And for the two numerical attributes in the dataset, the keyword “number” was implemented, which allows users to enter numerical values at their own discretion.

The third reason Gradio was such a great choice is simply because of run time. Although the model has to take in many attributes as input, it generates the output very quickly, which is a very important criterion for any UI.

As such, when accounting for its accessibility for users, how well-suited it was for the dataset, and the quick run time, Gradio was a great choice.

Conclusion and Discussion

Customer retention is very important for companies that want to maximize their profits. While attracting new customers spells success, it is equally important to keep current customers. The dataset used for the project contains multiple attributes to predict customer churn, most of which were categorical. After one-hot encoding the categorical variables and testing different models, it was determined that the best model was the logistic regression model, with an accuracy of about 78%.

There were certain criteria for determining which user interface library to implement. The Gradio library in Python was a good fit according to all of the criteria: user-accessibility, well-fit for the data, and quick run time. Since the logistic regression model was the best model, the UI was built based on that model, and it worked very well.

The biggest point of improvement lies in the data itself. The data cleaning process could have been smoother, but the biggest emphasis is the data itself. The chosen dataset was filled with categorical data, which meant that those variables had to be converted to numbers through one-hot encoding. Also, a lot of the attributes in the dataset had negligible correlation to customer churn, which led to those variables being omitted from the training and test sets. When considering the dataset, the model was more than satisfactory. As such, the project could have gone better if a better dataset with more correlated attributes was found.

References

Dataset:

https://www.kaggle.com/datasets/blastchar/telco-customer-churn

Literature Review:

Sabbeh, S. F. (n.d.). *Machine-learning techniques for customer retention: A comparative study - document view page*. Cluzters.ai.https://www.cluzters.ai/vault/279/2785/machine-learning-techniques-for-customer-retention-a-comparative-study?c=1597

Asthana, P. (2018). A comparison of machine learning techniques for customer churn prediction. https://acadpubl.eu/jsi/2018-119-10/articles/10b/2.pdf

Inzaugarat, E. (2021, April 28). *How machine learning can help with customer retention*. Medium. https://towardsdatascience.com/how-machine-learning-can-help-with-customer-retention-6b5bf654e822