Capstone Project

Customer Churn Prediction

By Ryan Stamp
14/04/2023

Table of Contents

Contents

Problem Statement

Business to Customer (B2C) and Business to Business (B2B) organisations are always at risk of losing current customers and forecasted future revenue. Known as churn, it occurs for multiple reasons, including cancellation of their product or service, switching to a competing provider, failing to renew their subscription or closing their account. Having involuntary customer turnover (or churn) leads to higher customer acquisition costs (CAC) (the cost of sales and marketing required to gain a new customer).

Having a higher CAC, leads to lost revenue in other areas of the business, as extra sales and marketing costs need to be allocated to gaining new customers to replace lost ones.

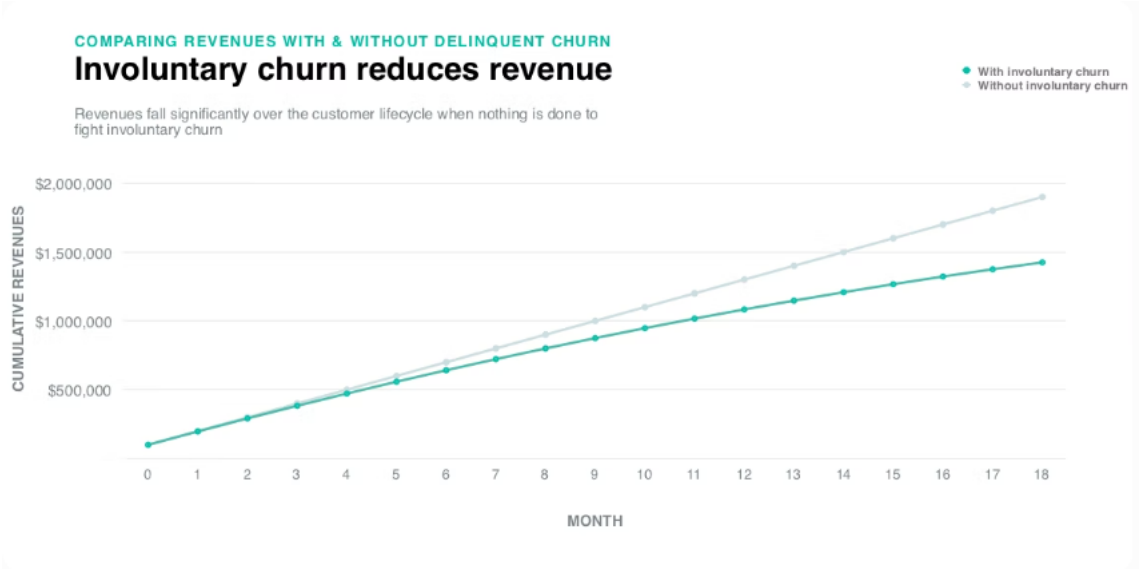


Fig.1 Example in revenue differences with and without involuntary churn

Assessing what variables are affecting customer churn helps to predict and either minimise or stop the future loss of customers.

There are several key performance indicators (or variables) which can determine the reason for customer churn, some of which include, levels of customer support and net promoter score (NPS), competitor pricing, the likelihood of customers renewing or upgrading their current service.

Customer segmentation and insights can also provide analysis of customer churn, including which demographic they fall into (age, gender and what they have purchased), as well as how engaged they are with your product/service and whether they are receiving the correct value for what they are paying.

If successful analysis is undertaken, negative churn is able to exist, where the compounding revenue from existing customers (renewed subscription or repeat business, upsells etc.) can offset the revenue lost from customer churn.

Industry/Domain

The organisation provides financial advice to its customers digitally, delivered in the form of electronic newsletters, webinars, and investment recommendations.

The digital newsletter can be subscribed to either on an annual or monthly basis. The company also provides support for customers to reach out to a via a contact centre that can help them with

product-related questions and signup/cancellation-related queries. The organisation has external competitors providing a similar product and service (other financial advice providers for example), as well as traditional financial institutions (banks).

Stakeholders

The key stakeholders for this project are the Sales and Marketing Manager and Senior Financial Advisor. They are looking to identify key areas of customer churn and projected monthly subscription revenue based on the last 5 years of sales.

Business Question

Customer churn can significantly affect forecasted revenue and throw off future planning. Can we accurately predict customer churn (and why customers leave)? The business would like to estimate whether the current data collected about its customers will accurately assist in predicting this.

Data Question

The data question will be 'can we predict the churn of future customers?'

Data

The dataset was sourced from Kaggle and is likely a sample set of customer and sales data. The databases provided would likely resemble real life data found within a typical company within the industry.

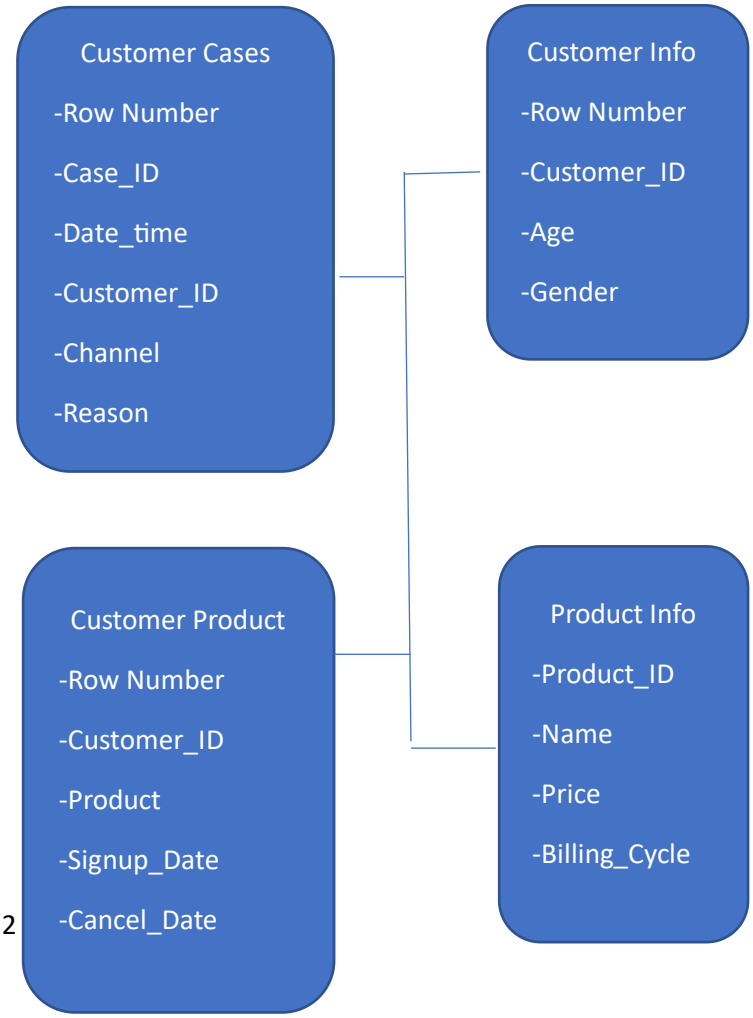


Fig. 2

Figure 2 shows the data structure (sourced from Kaggle). The data consists of 508932 records across 5 years. The data is available to download.

https://www.kaggle.com/datasets/gsagar12/dspp1

Data Science Process

Data Analysis

The dataset was four tables – customer cases (records of contact with customer support), customer info (customer demographics), customer product (subscription type, signup and cancellation dates) and product info (product name and details). Not all customers had used the customer support desk and so had no data for these variables. The tables were loaded in Jupyter notes as dataframes. The tables were then merged based on customer and product ID's.

Performed feature engineering on the final dataset. This included separating out signup year and month as well as cancelling year and month into new features, as well as categorising customer age in 4 age groups. Features were also changed from categorical and object (text) data types to integers (numbers) where necessary.

The target variable (churn) was created from the customer cancellation rate.

```
0    396447
1    112485
Name: chur, dtype: int64
```

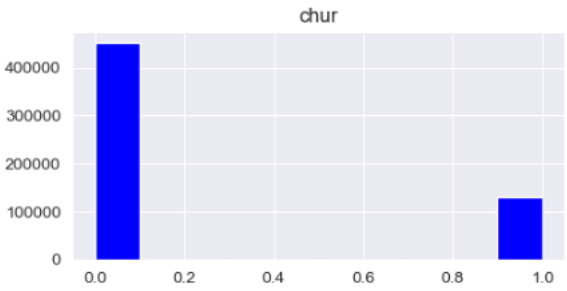


Fig.3 Customer cancellations (churn)

It's approximately 20% of the dataset (which is considered mildly imbalanced) however this was still scaled. The length of customer subscriptions ranged from 1 day to 1826 days (five years).

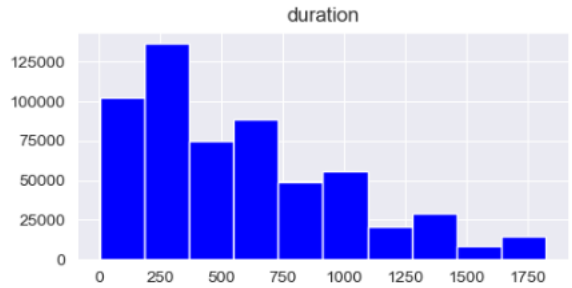


Figure 4. Customer subscription duration (days)

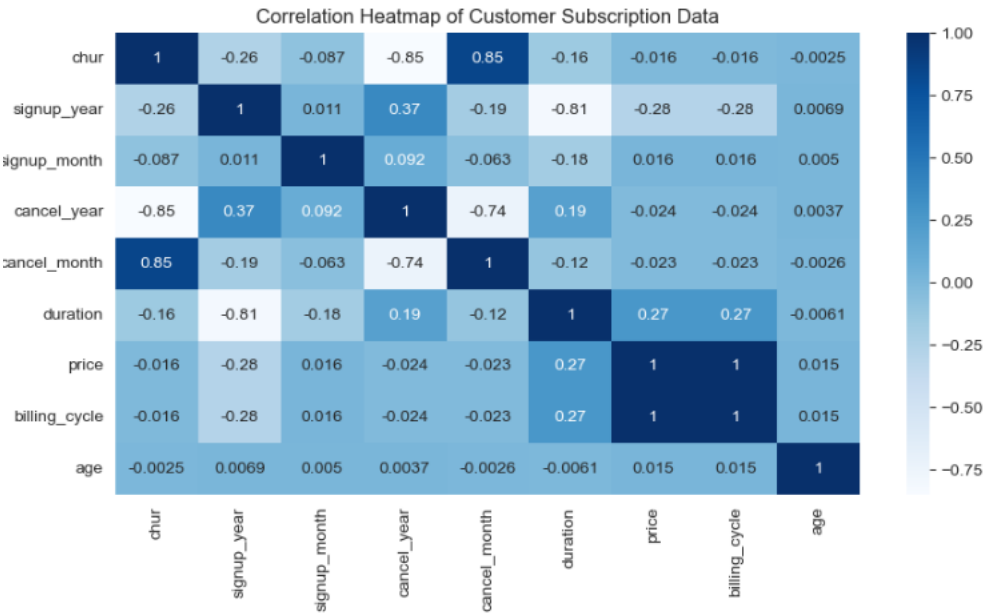
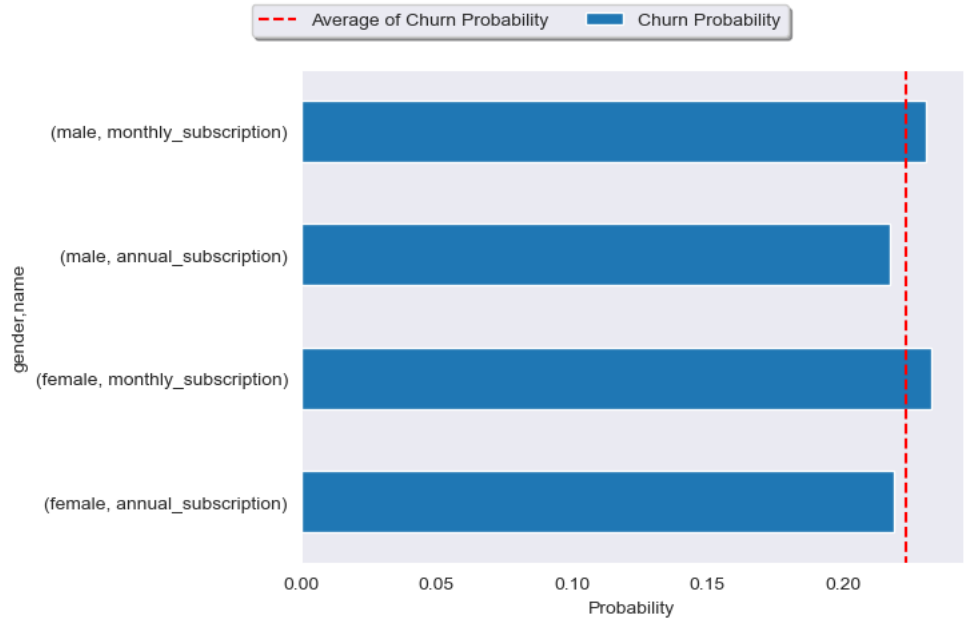


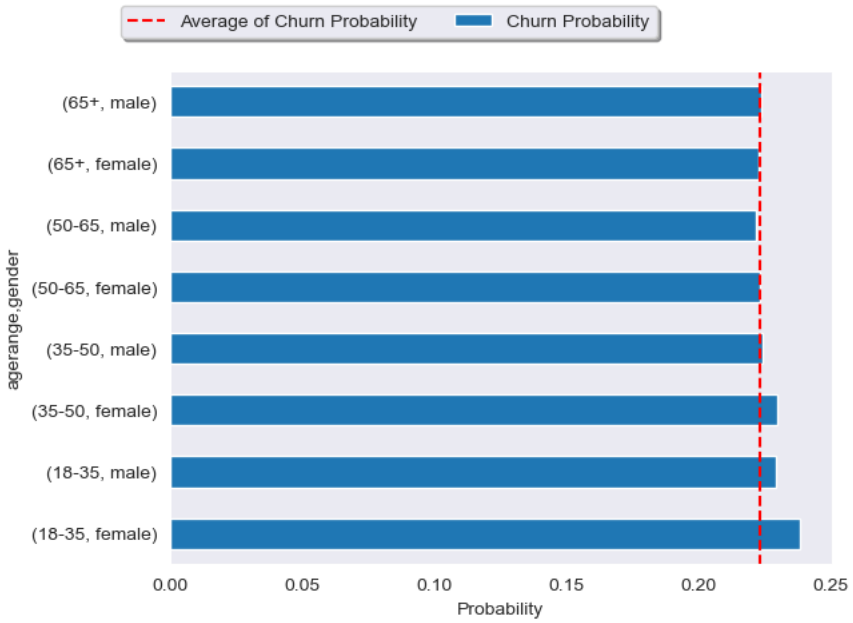
Figure 5. Plotted Pearson correlation heatmap to find out the correlation with the features.

Initial exploratory data analysis showed high correlation with the target variable yet this will understandably skew the modelling.

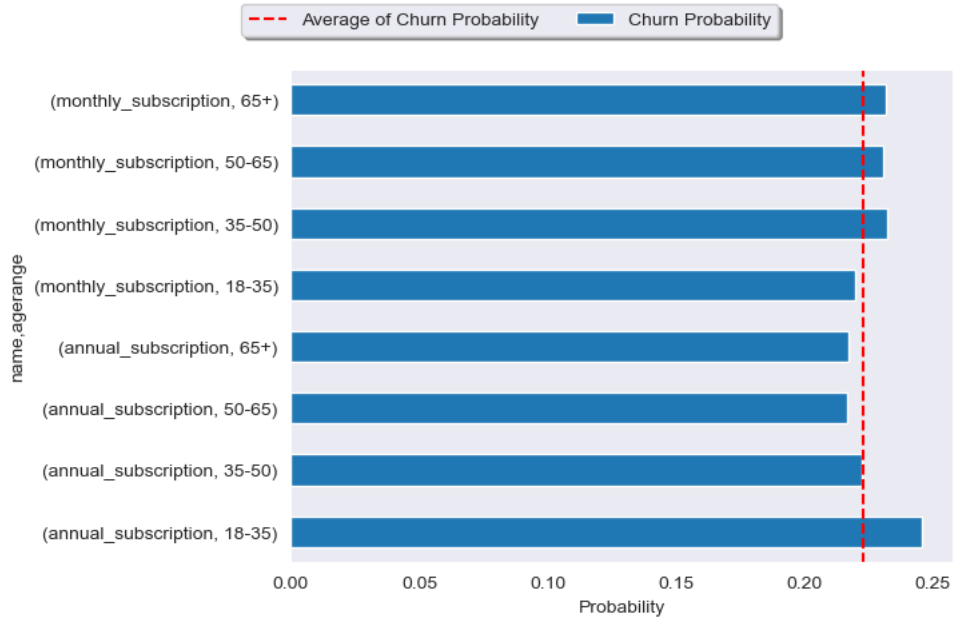
Plotted the probability of customer churn with variable pairs (product and customer segments).



Monthly subscribers have a higher probability of cancelling.



Females aged 18-35 have the highest probability of cancelling, followed by 18-35 year old males, then 35-50 year old females.



Annual subscribers aged 18-35 have the highest probability of cancelling, followed by monthly subscribers aged 35-50, then monthly subscribers aged 65+.

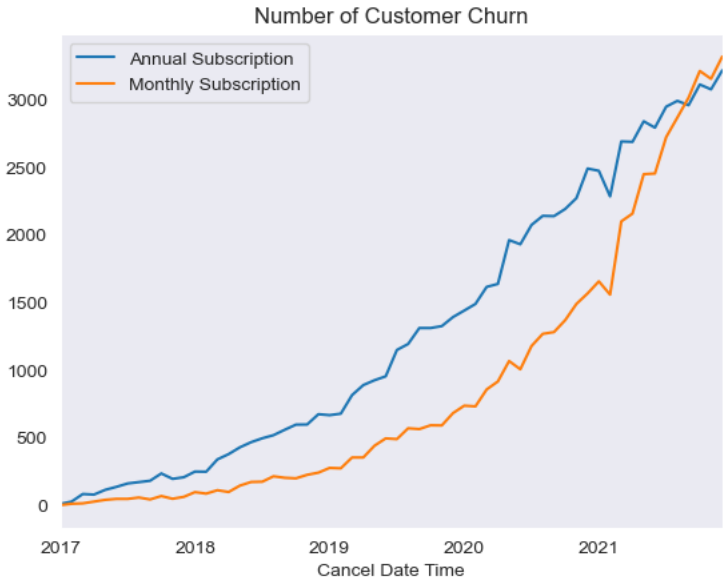


Fig.9 Cancellation Rate of annual and monthly subscribers

There was a higher total number of annual subscribers who cancelled until mid-2021, where a sharp spike at the start of 2021 saw the number of total monthly subscribers surpass this.

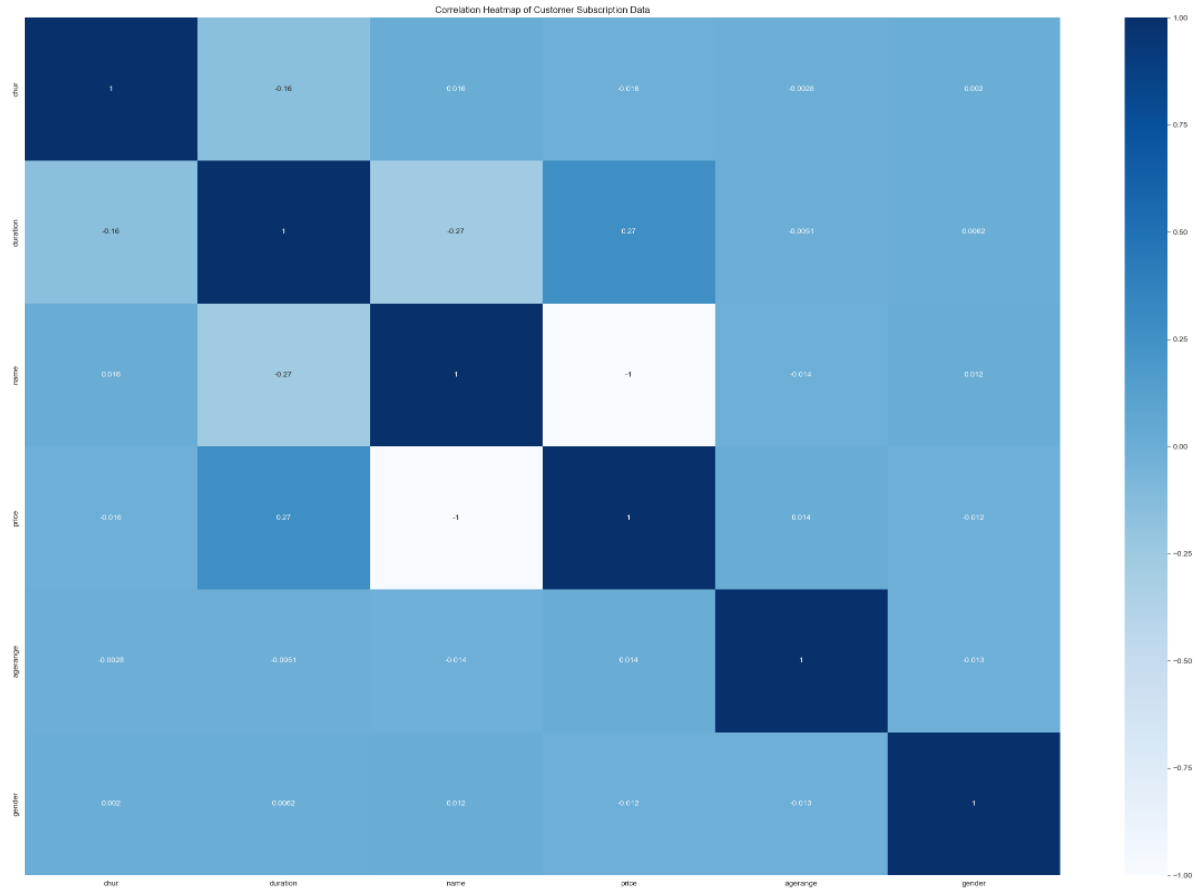


Fig. 10 Final Correlation heatmap (with engineered and removed features)

A second Pearson Correlation Heatmap shows more stable features to correlate with the target variable (churn). Duration is slightly stronger than the other variables in terms of correlation.

Data science process was completed within a Jupyter notebook. The exploratory data process is able to be reused with updated data in the future.

Modelling

The features selected were duration, (product) name, price, billing_cycle, agerange, gender, channel and reason. The target variable is chur(n). Some features are more strongly correlated with other features compared with the target variable. These were selected from the correlation map. Three models were used to predict the customer churn. The models were Multi-layer Perceptron classifier, Logistic Regression (unsuccessful) Naïve Bayes (also unsuccessful), Random Forest classifier and Decision Tree classifier . All models took similar time to run. Below figures show the each models performance metrics.

MLP Classifier

	precision	recall	f1-score	support
0	0.991013	0.787042	0.877328	170658
1	0.060977	0.659586	0.111634	3578
accuracy	0.784425	0.784425	0.784425	0
macro avg	0.525995	0.723314	0.494481	174236
weighted avg	0.971915	0.784425	0.861604	174236

Fig. 12 MLP Classifier Precision, Recall, F1 and Accuracy Scores.

	precision	recall	f1-score	support
0	1.000000	0.777870	0.875059	174236
1	0.000000	0.000000	0.000000	0
accuracy	0.777870	0.777870	0.777870	0
macro avg	0.500000	0.388935	0.437529	174236
weighted avg	1.000000	0.777870	0.875059	174236

Fig. 13 Logistic Regression Precision, Recall, F1 and Accuracy Scores.

	precision	recall	f1-score	support
0	1.000000	0.777870	0.875059	174236
1	0.000000	0.000000	0.000000	0
accuracy	0.777870	0.777870	0.777870	0
macro avg	0.500000	0.388935	0.437529	174236
weighted avg	1.000000	0.777870	0.875059	174236

Fig. 14 Naïve Bayes Precision, Recall, F1 and Accuracy Scores.

	precision	recall	f1-score	support
0	0.960494	0.791921	0.868100	164159
1	0.121631	0.469386	0.193199	10077
accuracy	0.773267	0.773267	0.773267	0
macro avg	0.541063	0.630653	0.530650	174236
weighted avg	0.911978	0.773267	0.829067	174236

Fig.16 Random Forest Classifier Precision, Recall, F1 and Accuracy Scores.

	precision	recall	f1-score	support
0	0.965696	0.790594	0.869416	165325
1	0.109751	0.478959	0.178581	8911
accuracy	0.774656	0.774656	0.774656	0
macro avg	0.537723	0.634776	0.523999	174236
weighted avg	0.921920	0.774656	0.834085	174236

Fig.16 Decision Classifier Precision, Recall, F1 and Accuracy Scores.

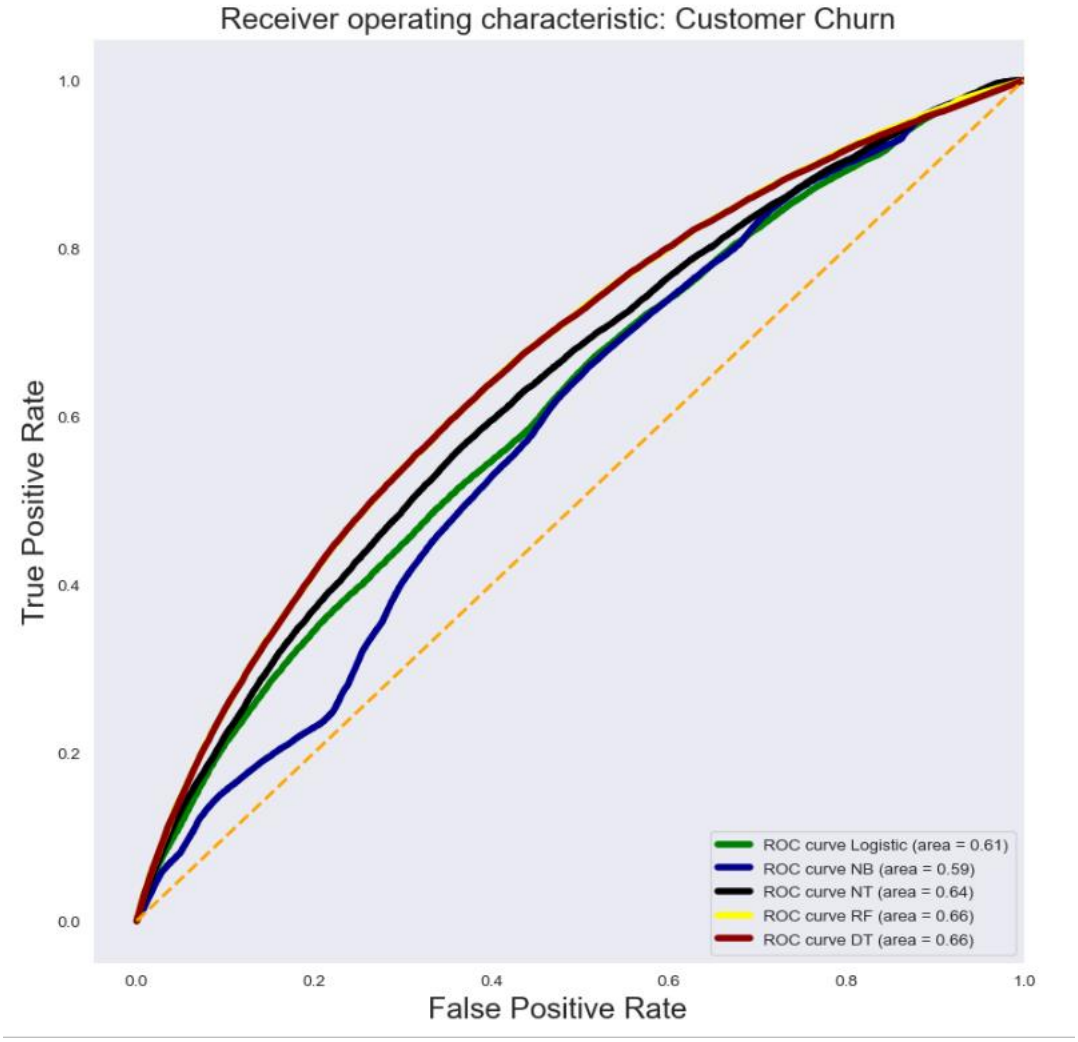


Fig. 17 ROC Curve comparing Logistic Regression, Naïve Bayes and MLP, Random Forest and Decision Tree Classifier.

MLP Classification model would be the preferred model as it has strongest recall of the target variable however the AUC result of .64 is not strong enough to recommend predicting future churn.

Outcomes

The EDA and process of feature engineering led to using the available features in the dataset. The Multi-Layer Perceptron Classifier model gave results with the highest accuracy. The results were not reliable enough to predict recommending however.

Implementation

From the results there would need to be further data collection to help achieve more reliable results from the models used. The models should then be re-tested to see if there is an increase in the AUC results (reliability).

Data Answer

The dataset does not have enough features to accurately predict customer churn reliably for future forecasting.

Business Answer

The business question was not answered satisfactorily. 78% is not strong enough accuracy nor is 66% reliability from the strongest model.

Response to stakeholders

Further data collection is required to achieve stronger results for future customer churn predictions. The project is ongoing before it can be fully deployed.

End-to-end solution

The business would need to plan for the correct infrastructure to maintain the current data frames and collection, as well as operating systems to run the model and display ongoing results.

References

https://towardsdatascience.com/how-to-balance-a-dataset-in-python-36dff9d12704

https://www.paddle.com/resources/customer-churn-analysis

https://developers.google.com/machine-learning/data-prep/construct/sampling-splitting/imbalanced-data#:~:text=A%20classification%20data%20set%20with,smaller%20proportion%20are%20minority%20classes.