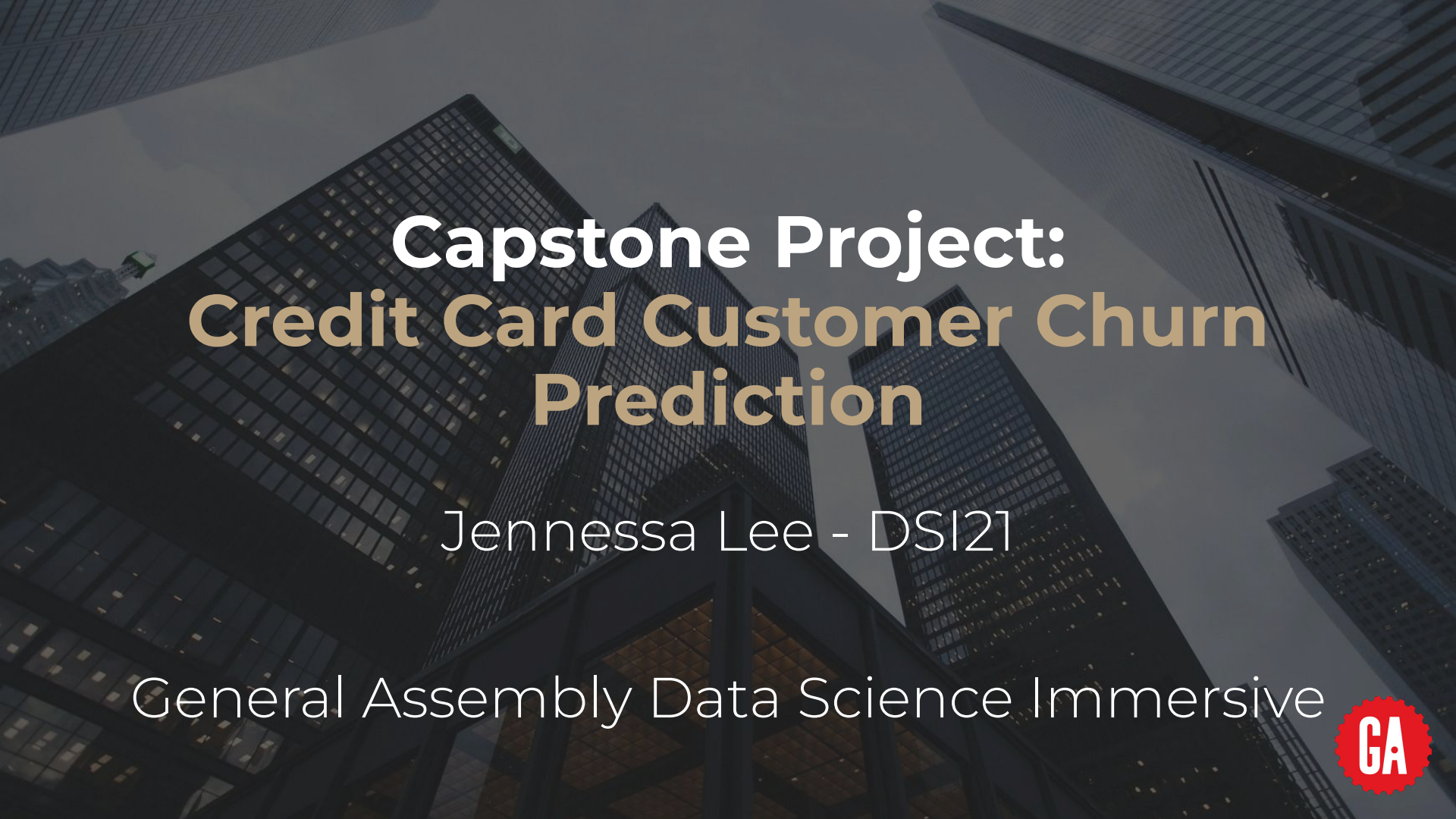
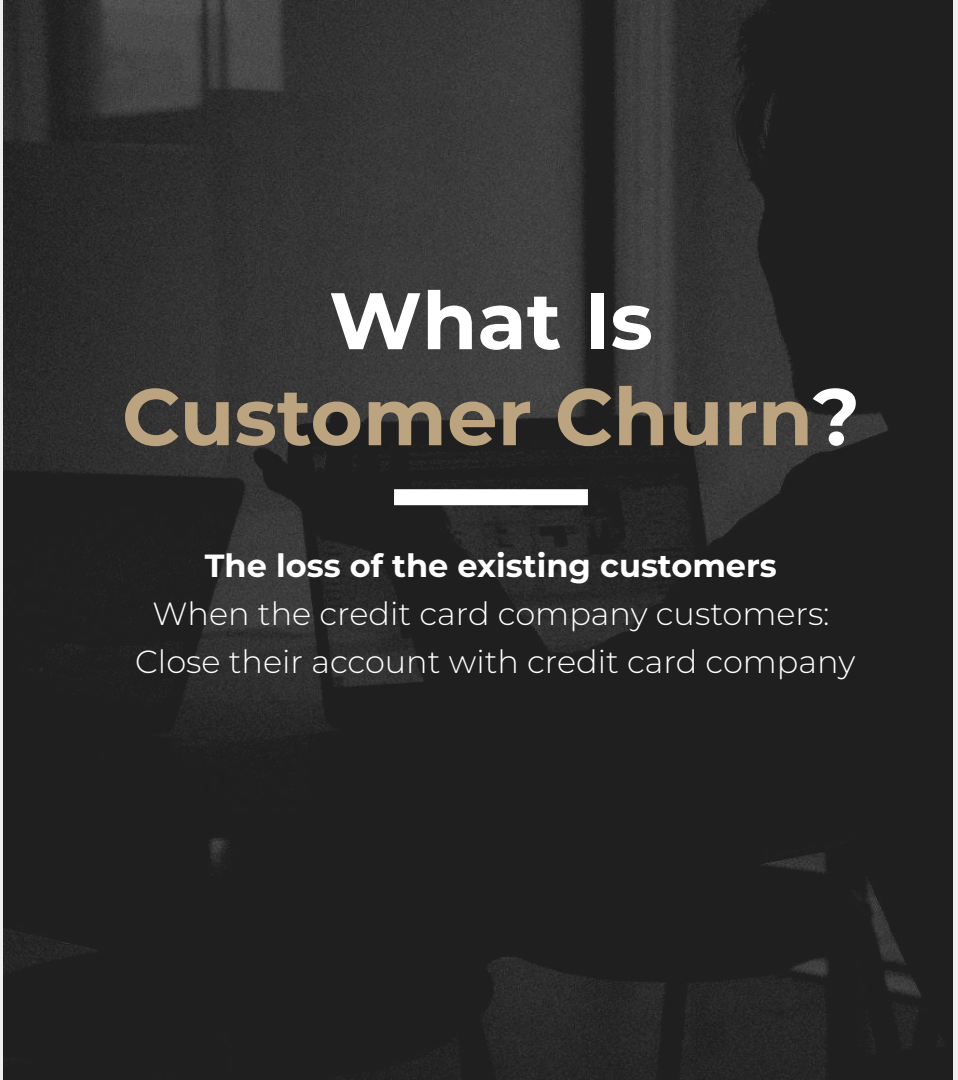


Capstone Project:
Credit Card Customer Churn Prediction

Jennessa Lee - DSI21

General Assembly Data Science Immersive



What Is Customer Churn?

The loss of the existing customers

When the credit card company customers:
Close their account with credit card company

What Is Customer Churn Rate?

Percentage of customers

Do not make another purchase using the credit card and terminate their relationship with the credit card company

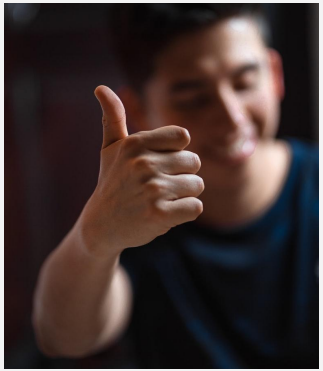
Capstone Project Process Flow

Business Problem

The Importance Of Customer Churn

Existing Customer Spend More

Existing customers spend **31% more** when compared to new customers & **50% more** likely to try new products



Loyalty Customer Is Profitable

⇧ 5% retention rates → **profits ⇧ up to 95%**

Retention Is Cheaper Than Acquisition

Acquiring a new customer can **cost 5 times** more than retaining an existing customer

To Be The Market Leader

Loyal frequent customer base → ⇧ market share → ⇧ corporation with more shops for promotions

Business Problem

Project Goal

Predicting for the credit card company that a certain customer is at a very high risk of churning

⇩

Credit card company can proactively provide a better product and service experience for retention

⇩

Reduce the credit card churn rate and successfully achieve customer retention

⇩

Success metrics: To achieve model accuracy that is better than the baseline

Hypothesis

Customer features are attribute for customer churn

1. If we **understand** what features are related
2. If we **proactively deal with** the related features

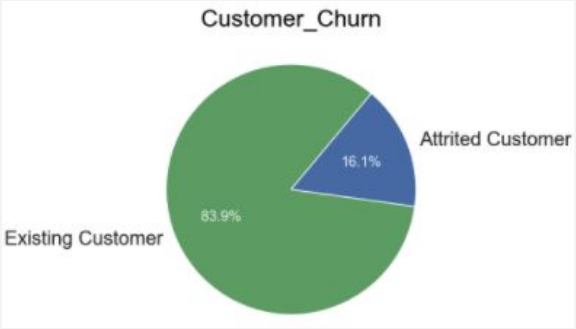
→ **Reduce the customer churn** for the company

About The Dataset

Data source: kaggle

- Consists of **10,127 total** no. of customers
- **1627 attrited customers** (16.07% out of all customers)
- **Strongly imbalance target variable**

→ Obstacle to train our model for predicting churning customers



About The Data

No Duplicates/ Missing Values:

```
# To check the number of rows and columns
print('Data Shape:', df.shape)

# To check if there is any duplicate data
print('Number of Duplicates:', len(df[df.duplicated()]))

# To check if there is any missing values
print('Number of Missing Values:', df.isnull().sum().sum())

# To check the column names
print(f"\nColumn Names:\n{df.columns}")
```

```
Data Shape: (10127, 23)
Number of Duplicates: 0
Number of Missing Values: 0
```

Unknown Values in 3 features:

```
# To check if there is any unknown values
print(f"\nUnknown Values %:\n{df.apply(lambda x: sum(x=='Unknown') / len(df))}")
```

```
Unknown Values %:

Education_Level
0.149995
Marital_Status
0.073961
Income_Category
0.109805
```

Data Cleaning

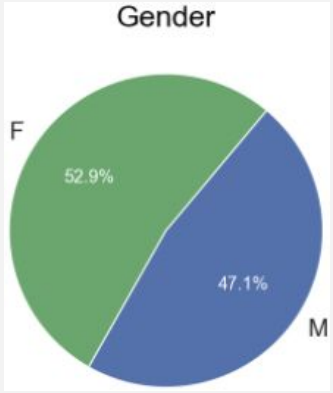
Drop Unwanted Columns:

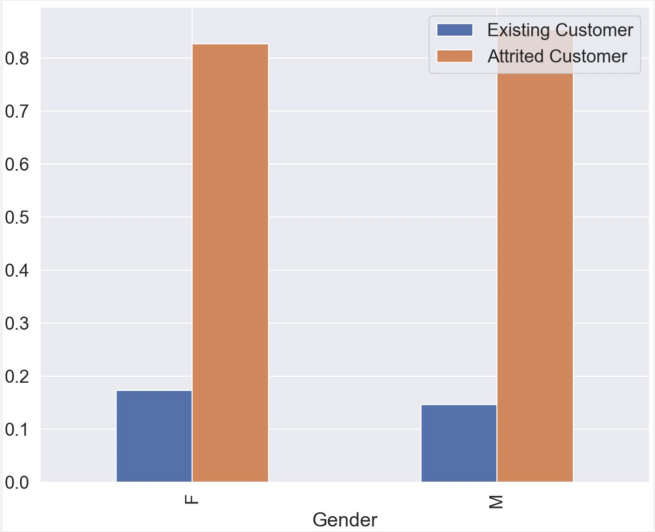
- Author suggests to **drop last 2 columns**
- Column with **unique client number** which has **no feature importance**

Rename & Standardise The Column Name:

- Replacing industry **jargon** column name with **straightforward** words

Exploratory Data Analysis - Gender



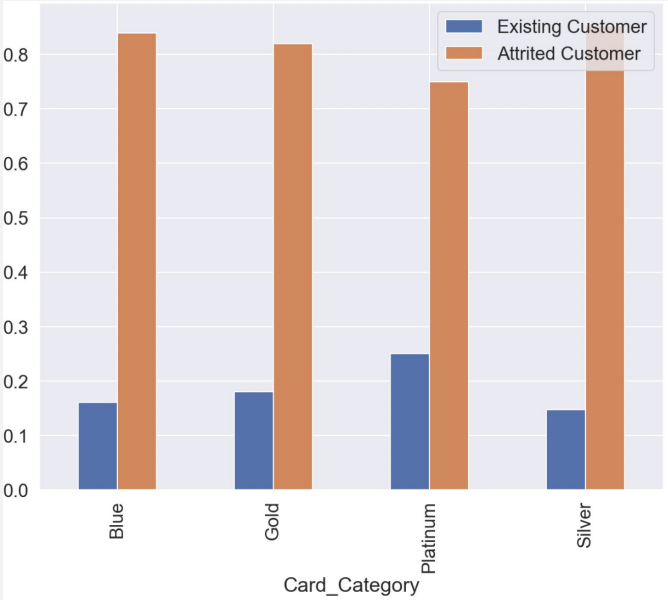


Proportion of gender is **balance**

No major difference between male & female for customer churn

Gender is **not a major factor** for customer churn

Exploratory Data Analysis - Card Category

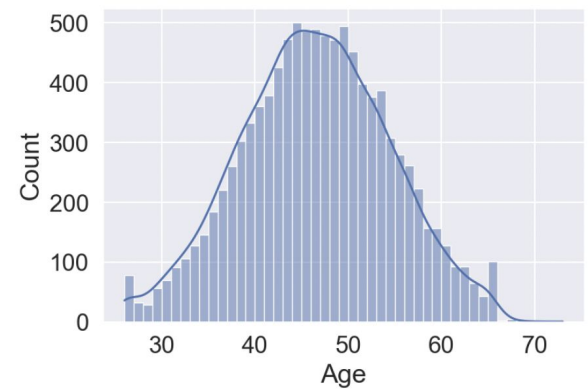


Majority card category = Blue Category (93.18%)

⇧ Card Category ⇩ No. of customer

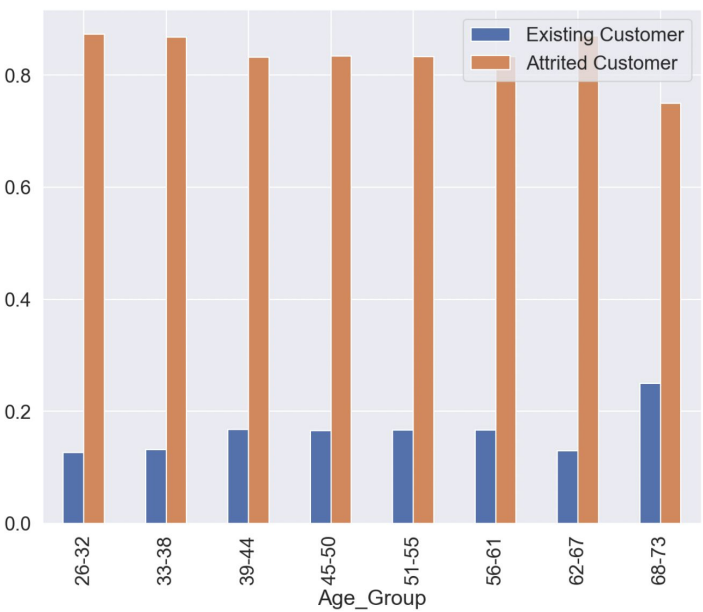
⇧ Card Category ⇩ Customer Churn

Exploratory Data Analysis - Age



Categorization

From continuous variable to categorical variable



Oldest age group (68 - 73): Lowest customer churn rate

It is believed that they may not know how to handle a credit card as they are old

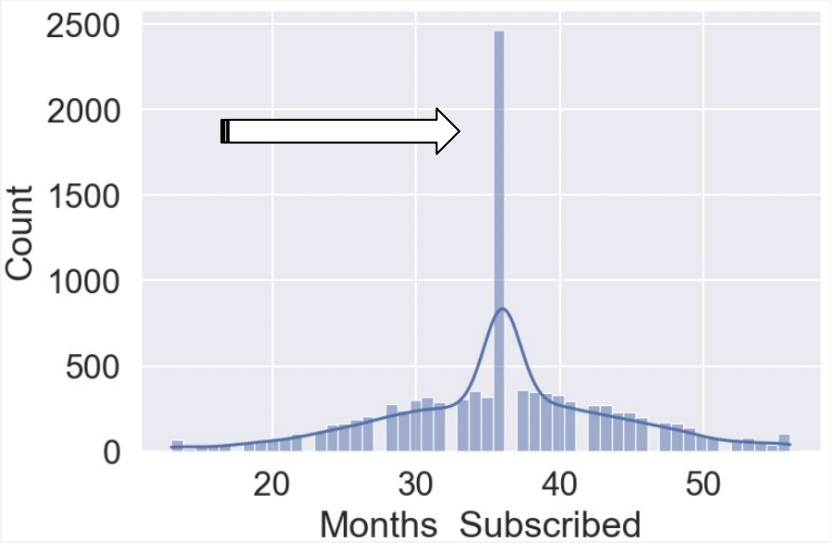
Exploratory Data Analysis - Months Subscribed

- Majority customers subscribed for 30 - 40 months

- **What Happened In 36 Months Ago?**

  Promotional campaign for attracting new customers to join the credit card service

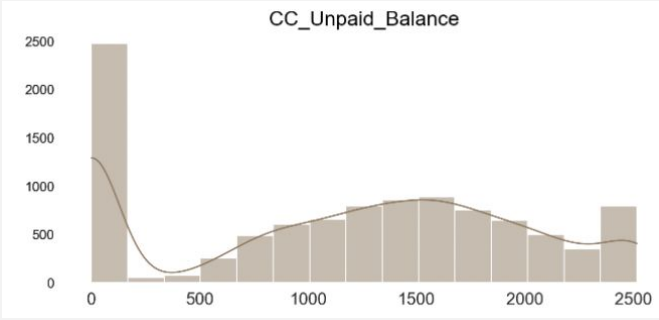
  Campaign is very successful as it is reflected by the data

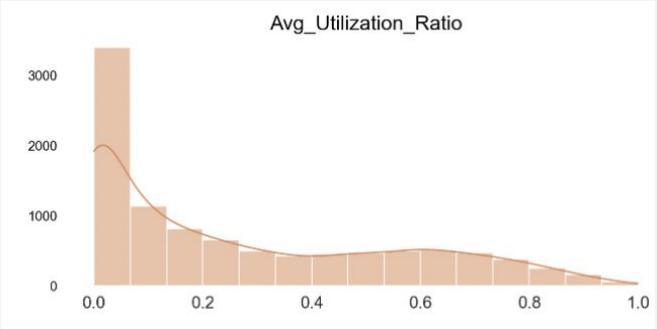


Exploratory Data Analysis - Columns With Zero Value

24.4% customers

Paid bill on time



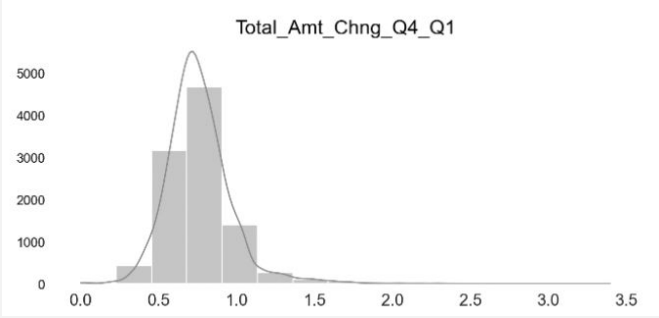


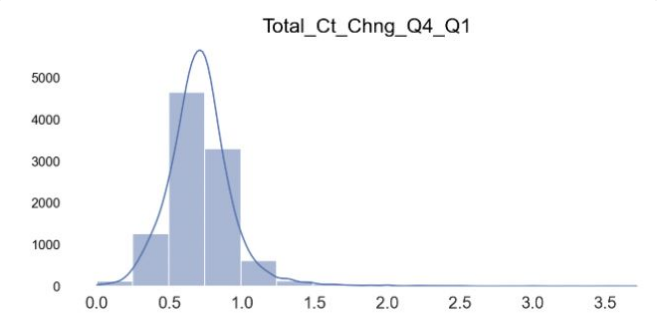
24.4% customers

Not using the credit card

0.05% customers

No change in total transaction amount

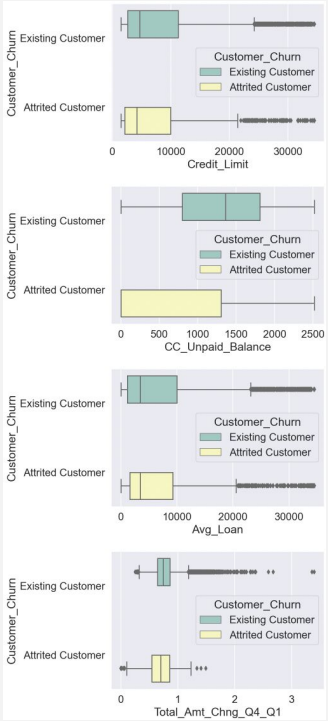




0.07% customers

No change in total transaction count

Exploratory Data Analysis - Numerical Features

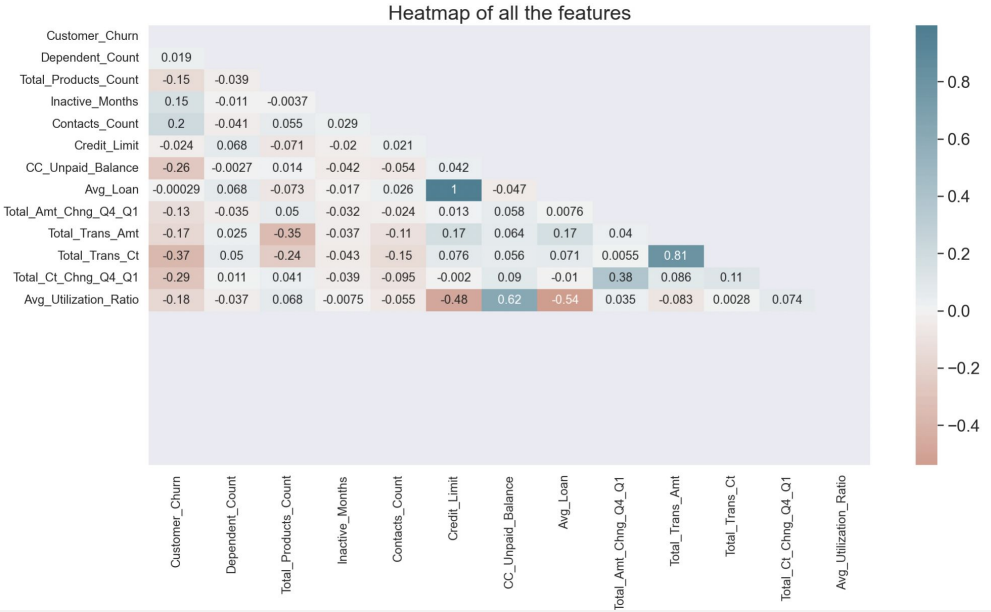


Will not drop the outliers at the moment

Reasons:

1. Size of dataset is not large
2. Will drop the row with unknown values in column of Education_Level, Marital_Status & Income_Category

Exploratory Data Analysis - Heatmap



- Avg_Loan and Credit_Limit have correlation coefficients 1
  → perfectly positive linear correlation
  → drop Avg_Loan column

- Total_Trans_Ct has the **highest positive linear correlation** with target variable, Customer_Churn

Exploratory Data Analysis - Target Variable

```
# Manually dummify the column to ensure Attrited Customer is 1 & Existing Customer is 0
df.Customer_Churn = df.Customer_Churn.replace({'Attrited Customer':1,'Existing Customer':0})
```

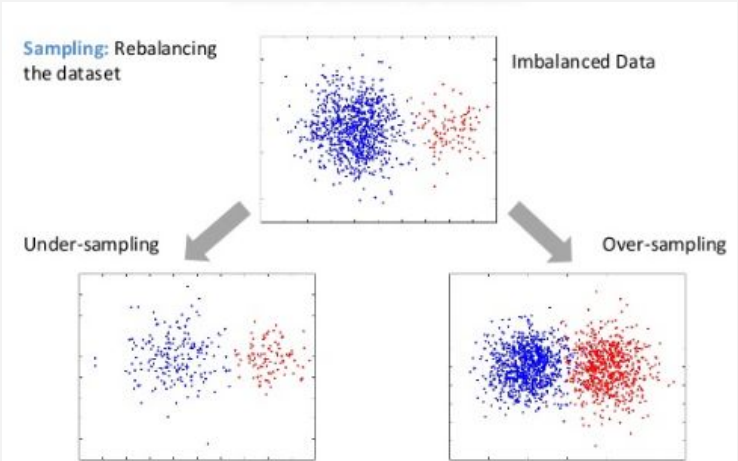
Manually dummify the target variable:

Attrited Customer = 1

Existing Customer = 0

Methodologies

- Modeling
  - 1. Without Sampling Technique
  - With Sampling Technique
    - 2. SMOTE For Oversampling
    - 3. NearMiss For Undersampling



Original Class Distribution:	**Balanced Class Distribution:**
Existing Customer (0): 0.83934	Existing Customer (0): 0.5
Attrited Customer (1): 0.16066	Attrited Customer (1): 0.5

Model With Best Performance

Model Comparison:

Modeling (SMOTE) > Modeling without sampling > Modeling (NearMiss)

Winner is **Modeling with SMOTE oversampling technique**

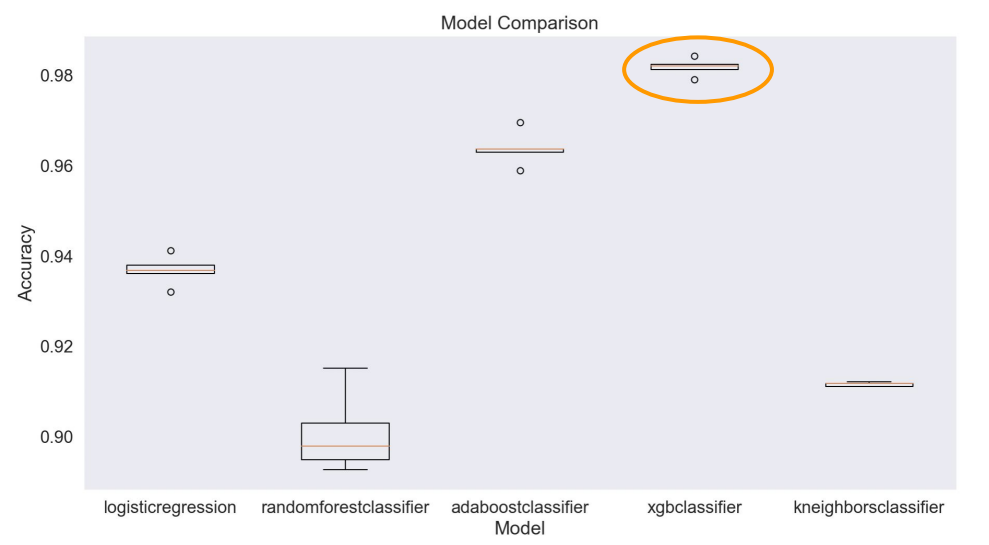
Best Machine Learning Algorithm for Method 2

Winner is

XGBoost

(eXtreme Gradient Boosting)

Only about 2% incorrect predicting churned customer



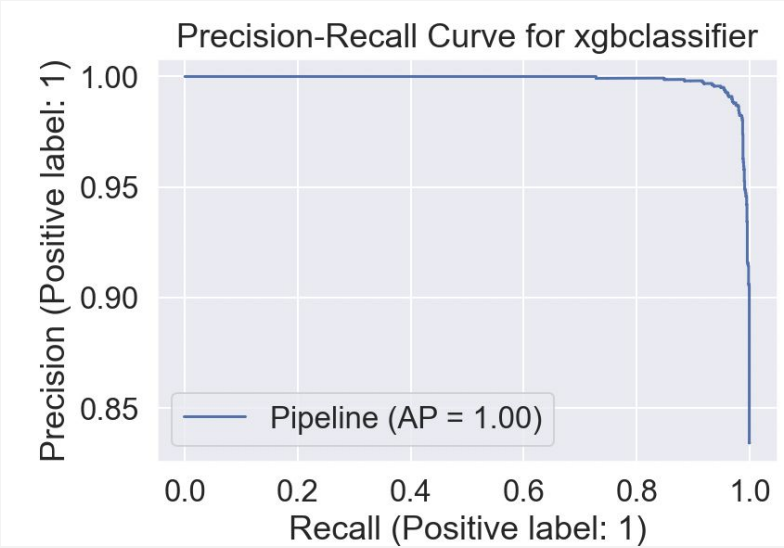
Model Performance

Precision score = **0.98**

Recall score = **0.98**

f1-score is **0.98**

Average cross-validation score = **0.98**



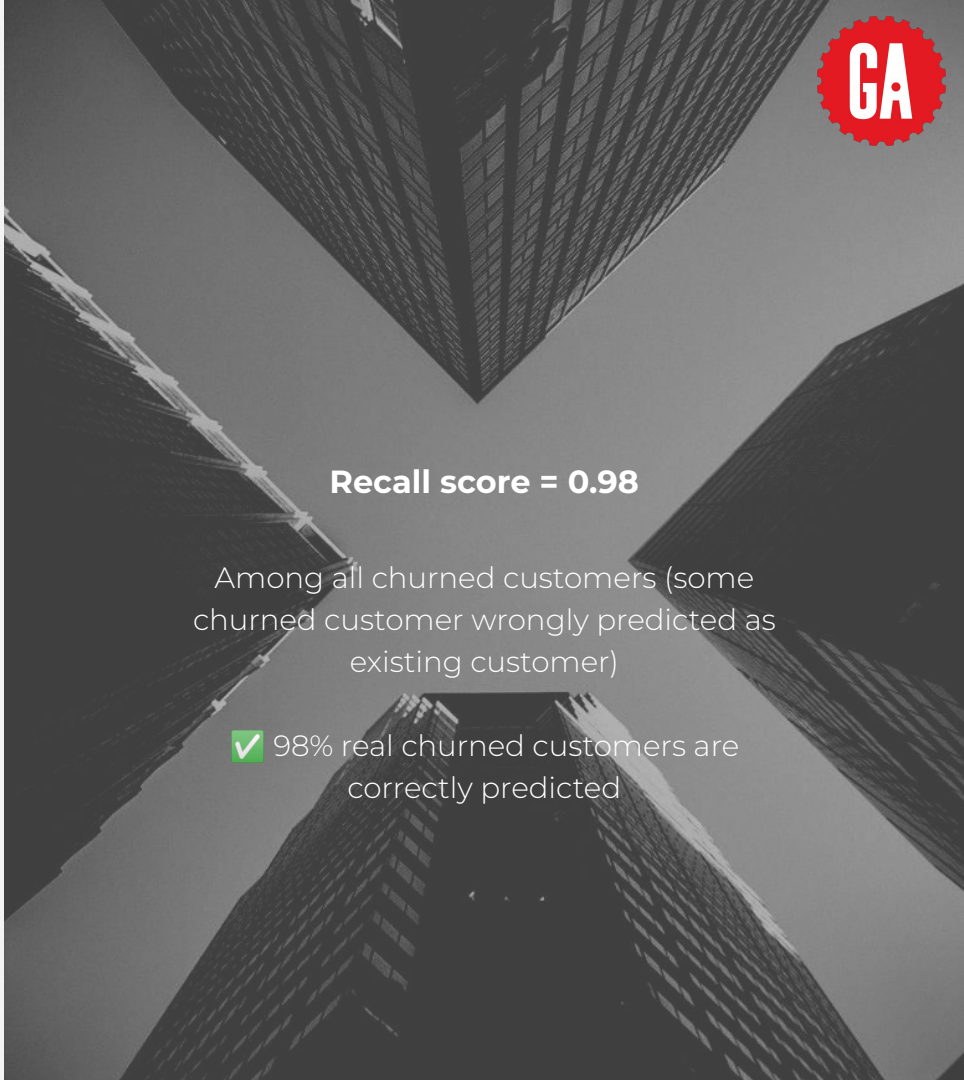
Credit Card Customer Churn Prediction

Summary

Precision score = 0.98

Among all predicted churned customers (some existing customer wrongly predicted as churned customer)

✅ 98% real churned customers are correctly predicted



Recall score = 0.98

Among all churned customers (some churned customer wrongly predicted as existing customer)

✅ 98% real churned customers are correctly predicted

Credit Card Customer Churn Prediction

Summary

f-1 score = 0.98

✅ A few false negatives
✅ Model is good at detecting customer churn
✅ Model is effectively assist the credit card company to notice those customers so that they can retain them

Limitations and Recommendation

Limitations

1. No external/ real-life data source due to business privacy issue

2. Very clean small-sized dataset → May not reflect the real-life situation

Recommendation

1. Build a recommender system:

   Gather other external datasets to establish user-item pair with rating for each credit card company products

   → Company can retain customers by offering exclusive promotions and offers

Thank You

Capstone Project: Credit Card Customer Churn Prediction

Jennessa Lee - DSI21

General Assembly Data Science Immersive