Bank Customer (Credit Card) Churn Prediction

This presentation provides an overview of predicting customer churn in the credit card industry. It covers the importance of identifying churn, the dataset used, and the steps involved in building a churn prediction model.

Understanding Customer Churn

- Customer churn refers to customers who stop using or leave services.
- Acquiring new customers is more expensive than retaining existing ones.
- Identifying churn is important to tailor marketing efforts and retain customers.

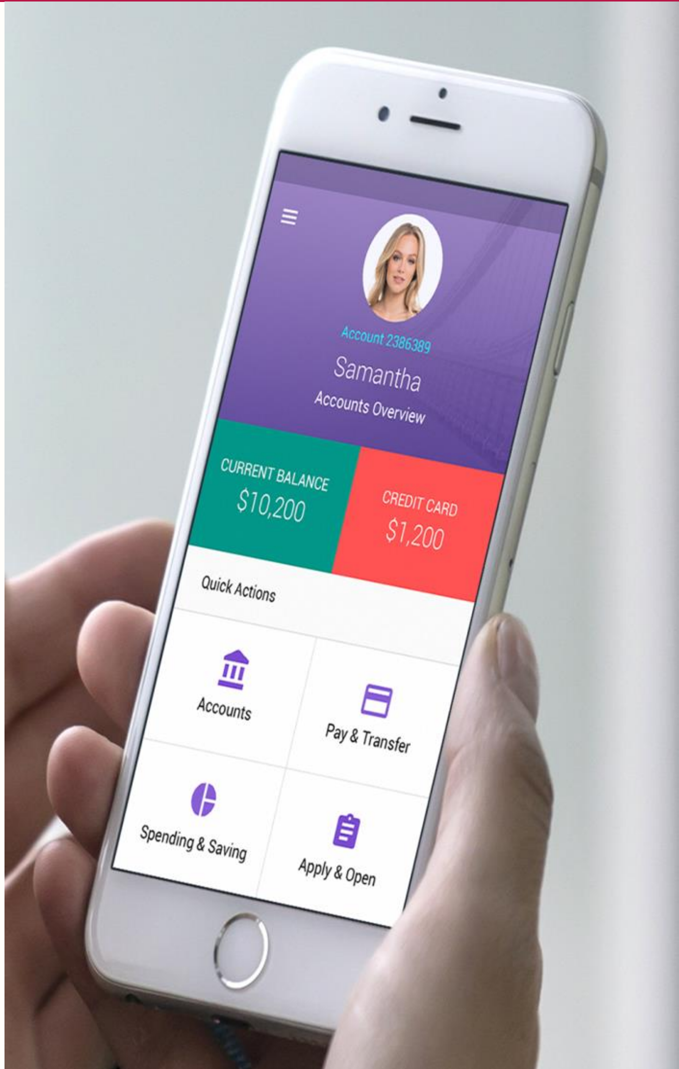
Photo by Pixabay

Problem Statement

- Analyzing customer demographics and numerical features can identify potential predictors of churn in the banking industry.
- Provide insights into potential predictors of churn.
- Combining demographic and numerical features can provide a comprehensive understanding of churn predictors and inform personalized retention strategies.

Features Selection

- Demographic factors such as gender, marital status, education level, and income category can reveal patterns related to churn.
- The dataset also includes data about each customer's relationship with the credit card provider, such as age, dependent count, number of months on book, and credit limit.



Data Preparation

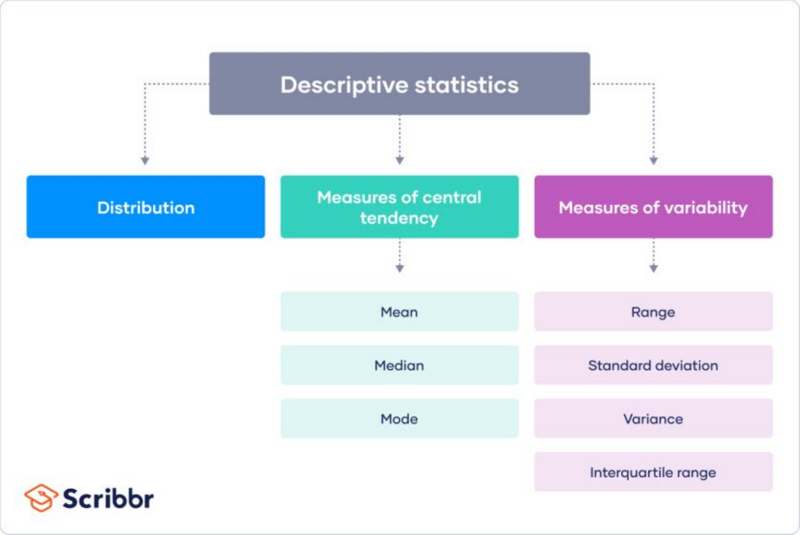
- Some variables may not be useful for the task at hand. In this case, variables such as Naive Bayes Classifier Attrition Flag, Total Amount Churned (Q1 to Q4), and Total Count Churned (Q1 to Q4) were identified as not useful and were removed.
- To incorporate gender as a predictor variable, a dummy variable was created. This involved assigning a value of 1 to male customers and 0 to female customers.
- To ensure data quality, customers with unknown marital status and education level were removed from the dataset.
  - Unknown values can introduce bias and uncertainty in models, potentially affecting the accuracy of churn predictions.

Photo by Pixabay

Exploratory Data Analysis

Descriptive Statistics

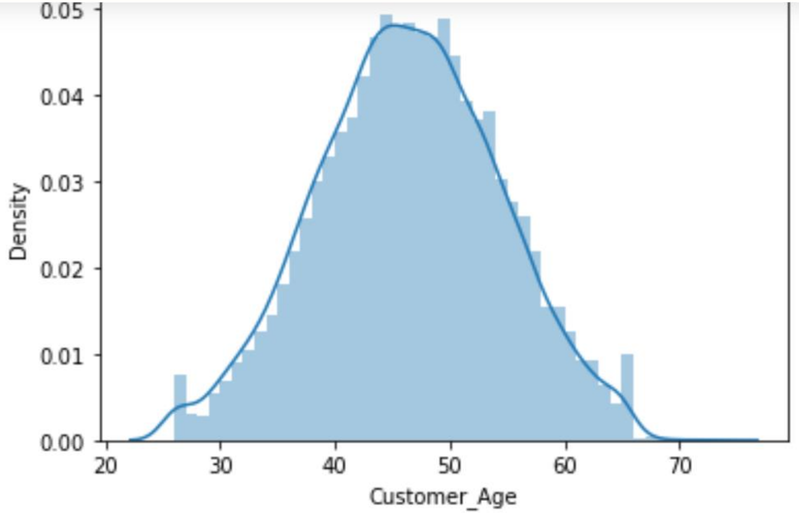
- Attrited customers had a lower average credit limit of $8,136.04 compared to existing customers with a higher average credit limit of $8,726.88. They also had an average of 36 months on books, while existing customers had an average of 35 months on book, suggesting that they remained with the bank for a longer period before their accounts were cancelled.
- Female customers were the predominant group in the dataset, representing 53% of the total observed customers.



Distribution Plot

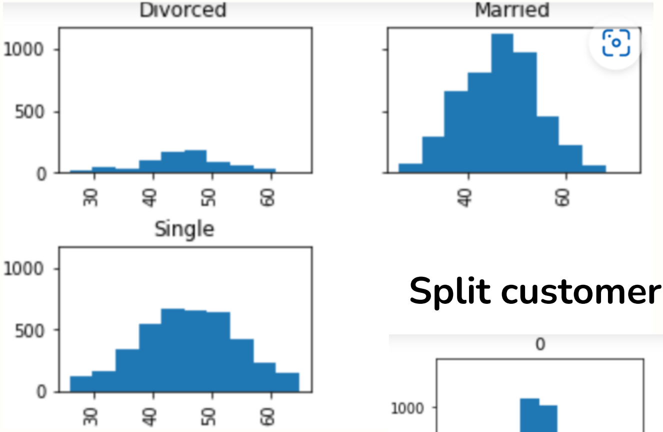
First let's use histograms to see the distribution of all the people

- As we can see here, the density distribution is relatively normal with the majority of customers around the ages of 40 to 50 years old.
- To split this distribution with categorical variables such as 'Married' status, we can use matplotlib.pyplot.hist(), which can be called with DataFrame object directly.

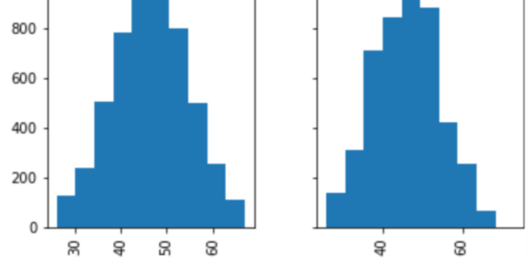


Categorical Distribution Plots

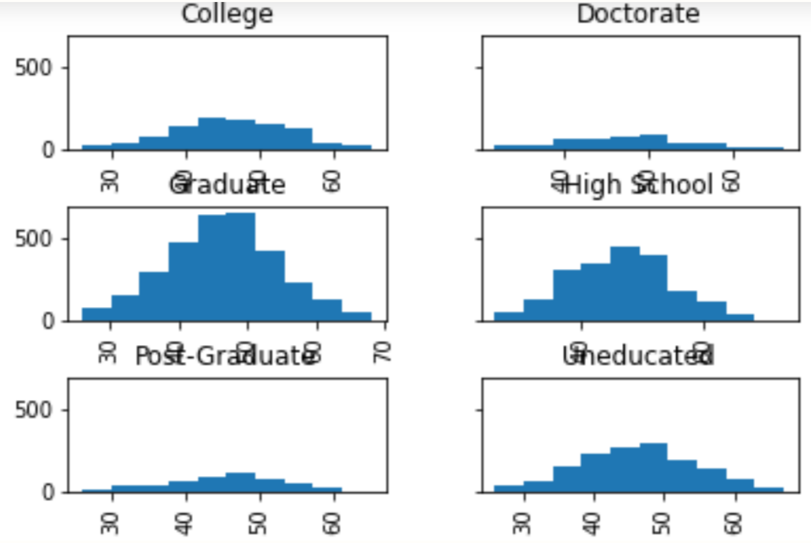
Split customers by ‘Marital_Status’



Split customers by ‘Gender’

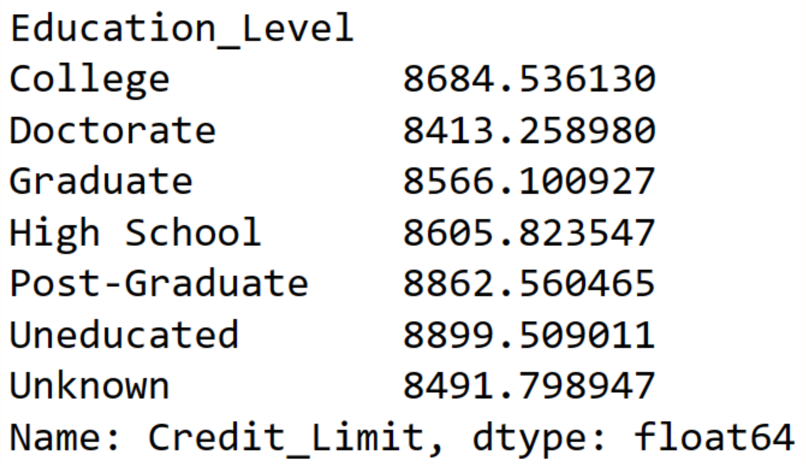


Split customers by ‘Education_Level’



Customer Relationship with Education Level

- To take a further look into one of the categorical variables, if we were to group customers by their education level place and compute their average credit limit.
- It appears that customers uneducated have the highest average credit limit compared to the other five education levels.
- Higher credit limits for middle-age customers are generally uneducated compared to the rest.

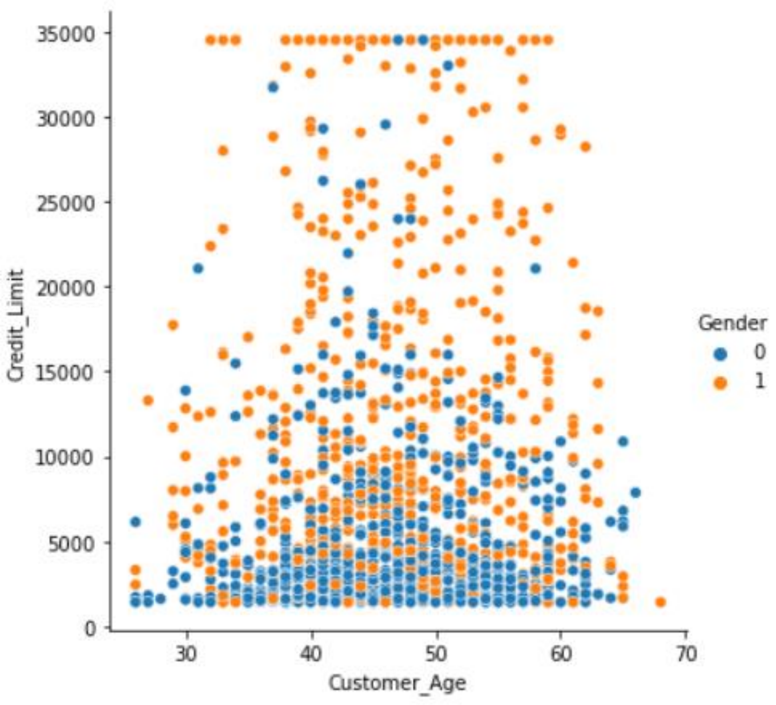


```
Education_Level
College            8684.536130
Doctorate          8413.258980
Graduate           8566.100927
High School        8605.823547
Post-Graduate      8862.560465
Uneducated         8899.509011
Unknown            8491.798947
Name: Credit_Limit, dtype: float64
```

Data Visualizations

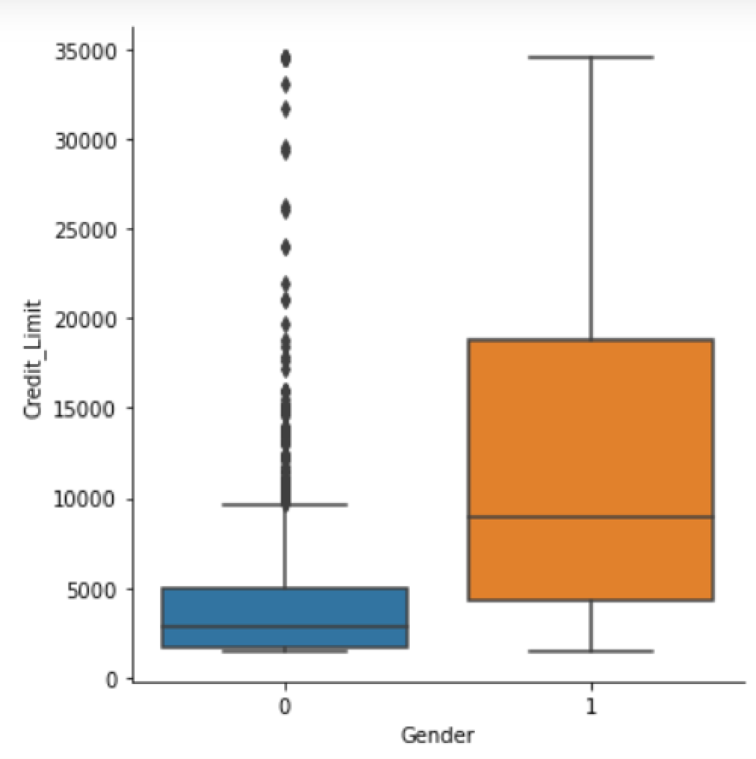
Relationship between Numerical Variables

- To illustrate the relationship between numerical variables, we also include the customer age and credit limit. In addition, there is also the inclusion of gender as a categorical variable.
- Using `seaborn.relplot` - show the relationship between customer age and credit limit variables. There is a slightly positive relationship with the majority of customers ages from 40 to 50 and credit limits from $3,000 to $5,000.



Relationship between Categorical Variables

- With the addition of a color argument of the categorical variable gender, we start to see that, on average, male customers have higher credit limits than female customers.
- Using `seaborn.catplot` - plot with categorical data gender with the distribution of credit limit on two boxplots (male and female). We can confirm that male customers have a higher average credit score than female customers. The distribution of female customer credit limits are negatively skewed where $10,000+ credit limit skewed the distribution.



Building Machine Learning Model

Customer Churn Prediction Model(s)

Split Training and Testing Set

- Feature variables includes all the numerical and categorical variables except for the `Attrition_Flag`.
- Target includes the categorical variable `Attrition_Flag` whose values are made up of either `Existing Customer` and `Attrited Customer`.
- Split data into training and testing set with the testing set size equal to 0.3, or 30%.
- Training set contains 7088 observations while Testing set contains 3039 observations.

Logistic Regression

Choose model:
`sklearn.linear_model.LogisticRegression`

1. Convert the Categorical data to Numeric features using `OneHotEncoder`
2. Standardize the numerical features splitting the original data into 2 subgroups.
3. Apply these transformations on the training and testing sets
4. Use our `Logistic Regression` classifier as a model

Encoding Categorical Features

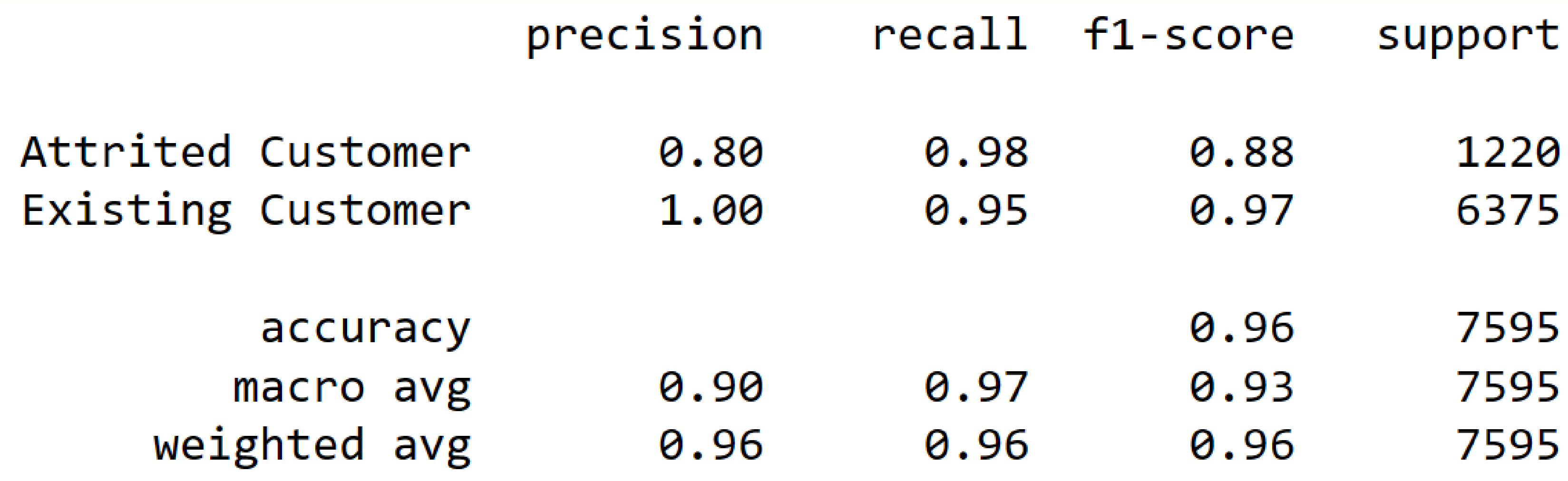
- The simplest way is to one-hot encode each categorical feature with the `OneHotEncoder`.
- We also want to standardize the numerical features. Thus, we need to split the original data into 2 subgroups and apply a different preprocessing: (i) one-hot encoding for the categorical data and (ii) standard scaling for the numerical data.
- The logistic regression predicts for attrited customers with an accuracy score of 84%.

Support Vector Machine

Choose model: `sklearn.svm.LinearSVC`

1. Reset model hyperparameter
2. Fit training and testing set with the SVC model
3. Print Classification Reports using the testing set and prediction of the training set
   a. Precision Score
   b. Recall Score
   c. F1-score Score
   d. Accuracy

SVM Classification Report



```
                   precision    recall  f1-score   support

Attrited Customer       0.80      0.98      0.88      1220
Existing Customer       1.00      0.95      0.97      6375

         accuracy                           0.96      7595
        macro avg       0.90      0.97      0.93      7595
     weighted avg       0.96      0.96      0.96      7595
```

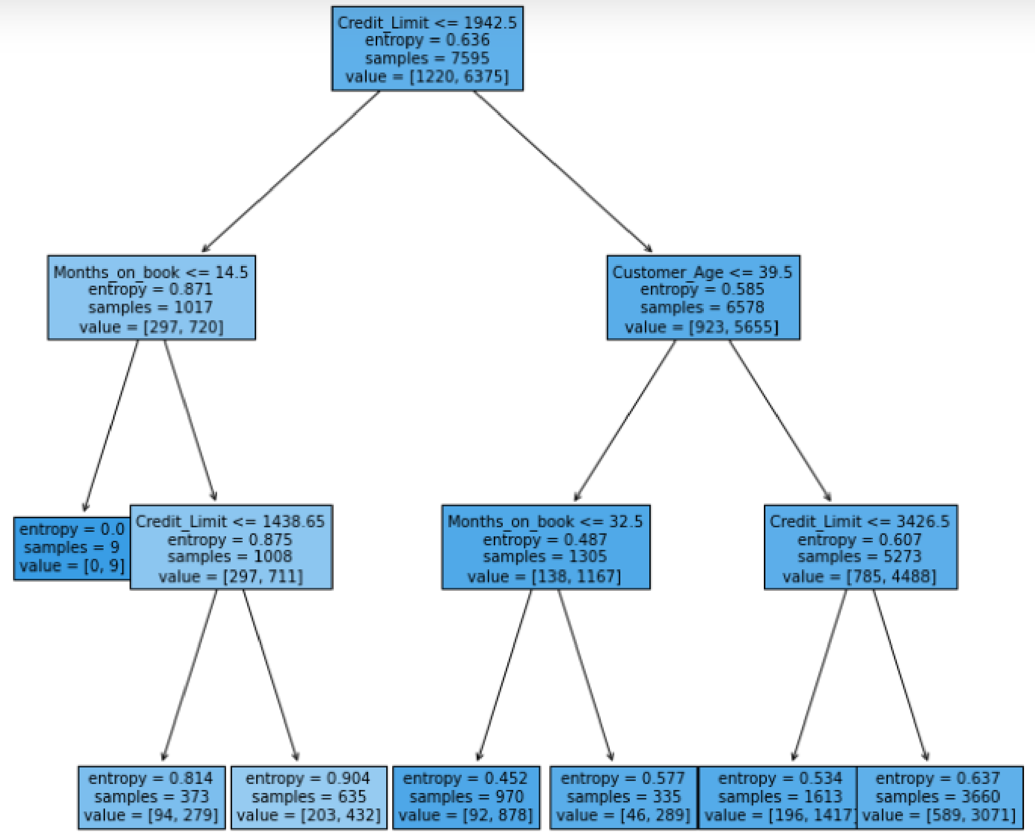
Kernelized SVM

Choose model:
`sklearn.preprocessing.StandardScaler`

1. Scaling the training and testing set using Standard Scalar method
2. Create a SVM Classifier using a linear kernel
3. Training the model using training set and predict the churn for test dataset
4. Calculate accuracy scores of training and testing set with an accuracy score of 0.8393, or 83.93%.

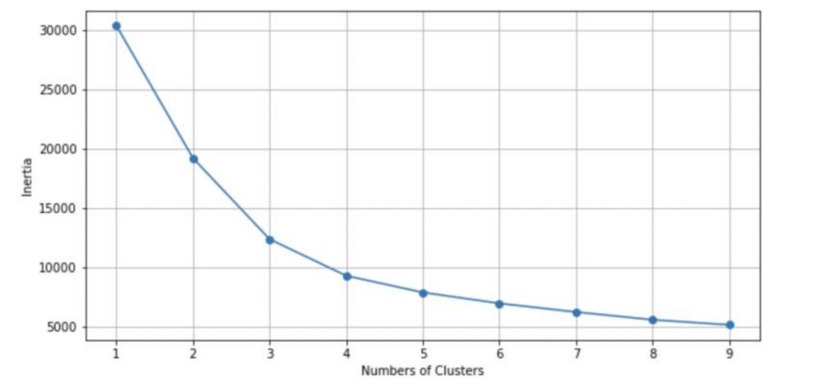
Decision Tree

- Classification tree answers a sequence of questions with each involving one of the numerical variables provided.
- In the image to the right, the tree has a maximum depth of 3 (may lead to overfitting on the training set).
- Predict the churn for test dataset results in an accuracy test score of 0.8393. Or 83.93%.



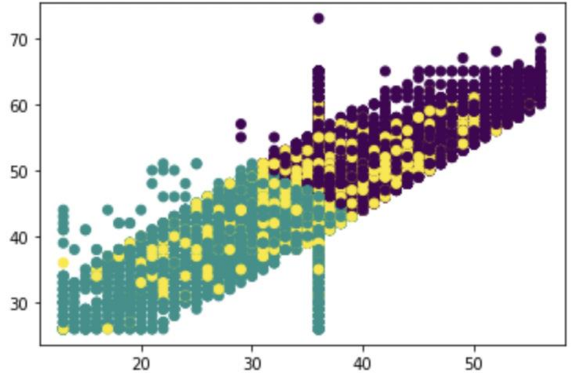
K Means Clustering

- Viewing the relationship between Customer Age, Months on Book and Credit Limit
- Important in understanding the how the variance in this data by person tends to group together
- Cluster to have 3 clusters



K Means Clustering Visualizations of Relationships

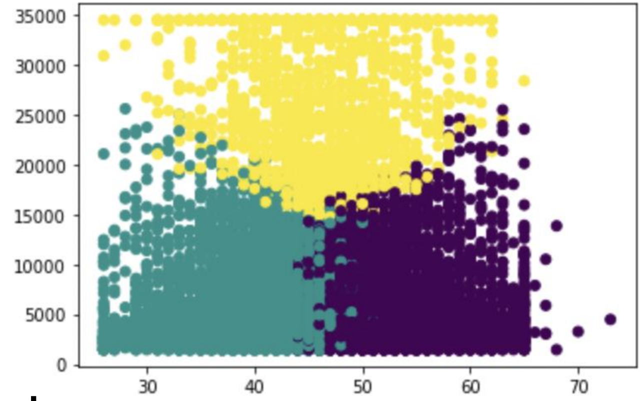
Months on Book x Customer Age



Customer Age

Months on Book

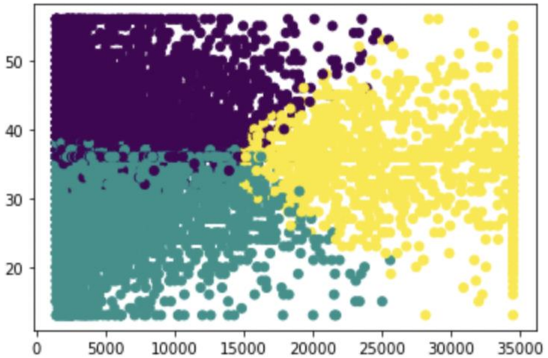
Customer Age x Credit Limit



Credit Limit

Customer Age

Credit Limit x Months on Book



Months on Book

Credit Limit

Findings of Predicting Customer Churn

1. The study explored the relationship between customers and numerical and categorical features such as Customer Age, Credit Limit, Marital Status, and Education Level. The analysis revealed that middle-aged, uneducated customers had the highest average credit limit.

1. The study predicted customer churn from the testing set using three classification models (Logistic Regression, SVM, and Decision Tree), and all models had an accuracy score of approximately 84%.

1. Clustering was used to investigate the relationship between customer age, months on book, and credit limit. The analysis determined the extent to which each pair of variables clustered together, providing insight into the relationship between these factors and customer behavior.

Photo by Pixabay

Thank you. Please feel free to ask any questions. 😄