PREDICTING CREDIT CARD CUSTOMER SEGMENTATION

Presented by:

Tadiparthy Dheeraj kumar AP21110010022

Reethu Bhargavi Sajjala AP21110010045

Beecha Venkata Naga Hareesh AP21110010839

Lahari Kotapti AP21110011172

Agenda

- INTRODUCTION
- BINNING
- DECISION TREE
- NAIVE BAYES
- RESULTS
- CONCLUSION

INTRODUCTION

This project proposes a approach for predicting credit card customer segmentation by combining binning techniques, decision tree analysis, and Naive Bayes classification. The methodology involves preprocessing the dataset using binning to simplify and enhance data analysis. A decision tree is then employed to create a transparent and interpretable model for customer segmentation based on key features. Additionally, Naive Bayes classification is integrated to improve predictive accuracy. The research, conducted on a real-world credit card dataset, demonstrates the effectiveness of the combined approach in accurately segmenting customers. The findings contribute to advancing credit card customer segmentation methodologies, providing valuable insights for targeted marketing and risk management in the financial industry.

BINNING

- Binning is a data pre-processing technique used to reduce the effects of minor observation errors.
- The original data values are divided into small intervals known as bins, and then they are replaced by a general value calculated for that bin. This has a soothing effect on the input data and may also reduce the chances of over fitting in the case of small datasets.

Why binning is used?

- Binning or discretization is used to transform a continuous or numerical variable into a categorical feature.
- It can also be used to identify missing values or outliers.

W**hat is the purpose of binning data?**

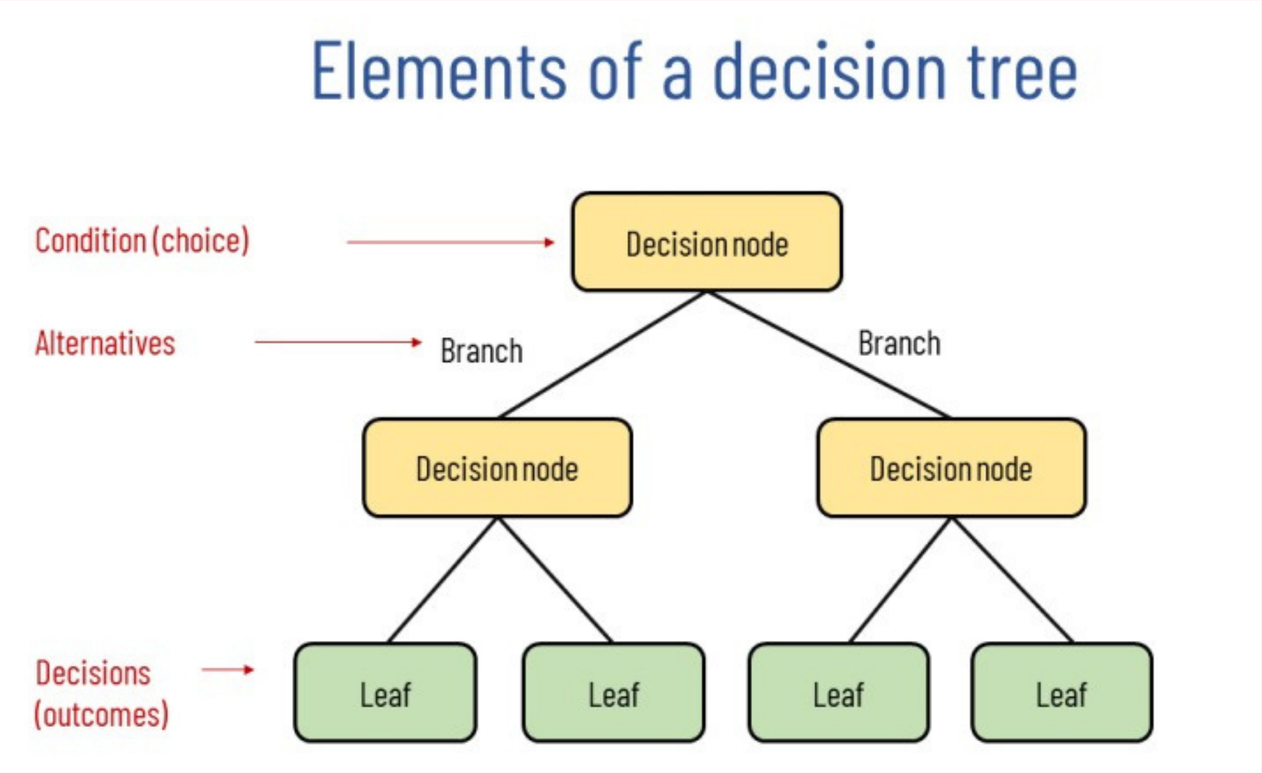
- Binning, also called discretization, is a technique for reducing continuous and discrete data cardinality.
- Binning groups related values together in bins to reduce the number of distinct values.

There are two methods of dividing data into bins and binning data:

- **Equal Frequency Binning:** Bins have an equal frequency.
- **Equal Width Binning:** Bins have equal width with a range of each bin are defined as [min + w], [min + 2w] …. [min + nw] where w = (max – min) / (no of bins).

DECISION TREE

- A decision tree is a type of supervised machine learning used to categorize or make predictions based on how a previous set of questions were answered.
- They are constructed using two kinds of elements: nodes and branches. At each node, one of the features of our data is evaluated in order to split the observations in the training process or to make an specific data point follow a certain path when making a prediction.



FORMULAE:

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

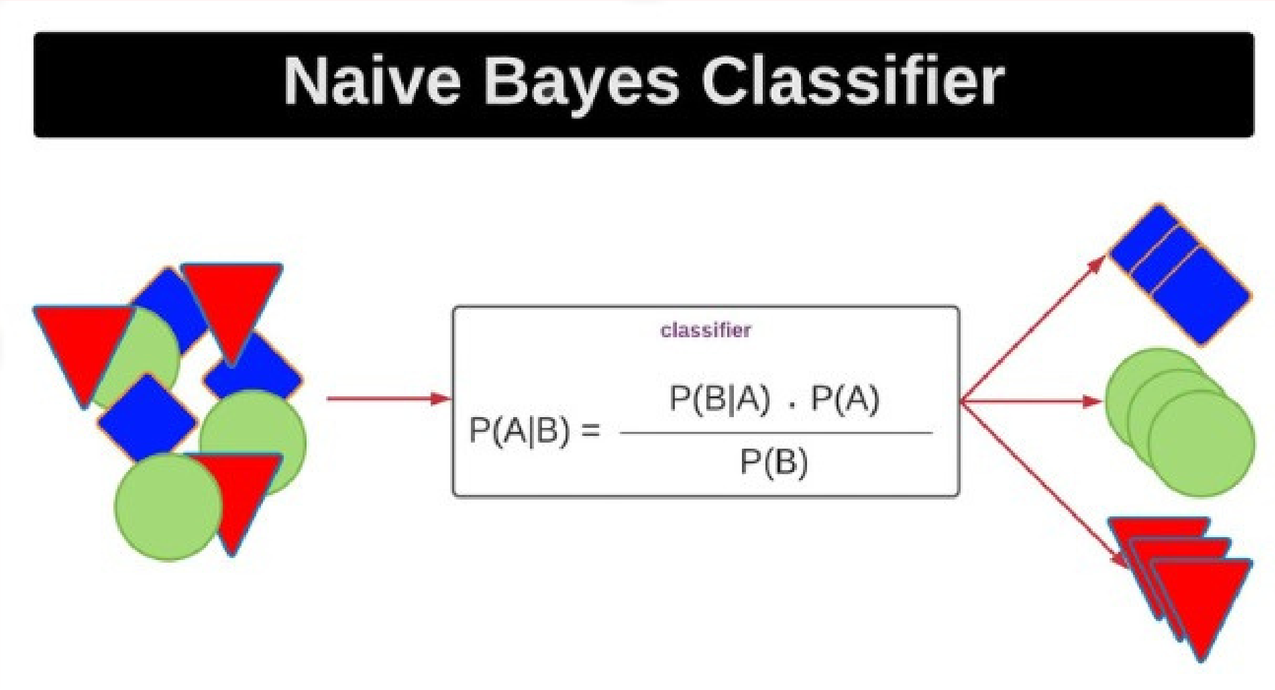
$$Gain(T, X) = Entropy(T) - Entropy(T, X)$$

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

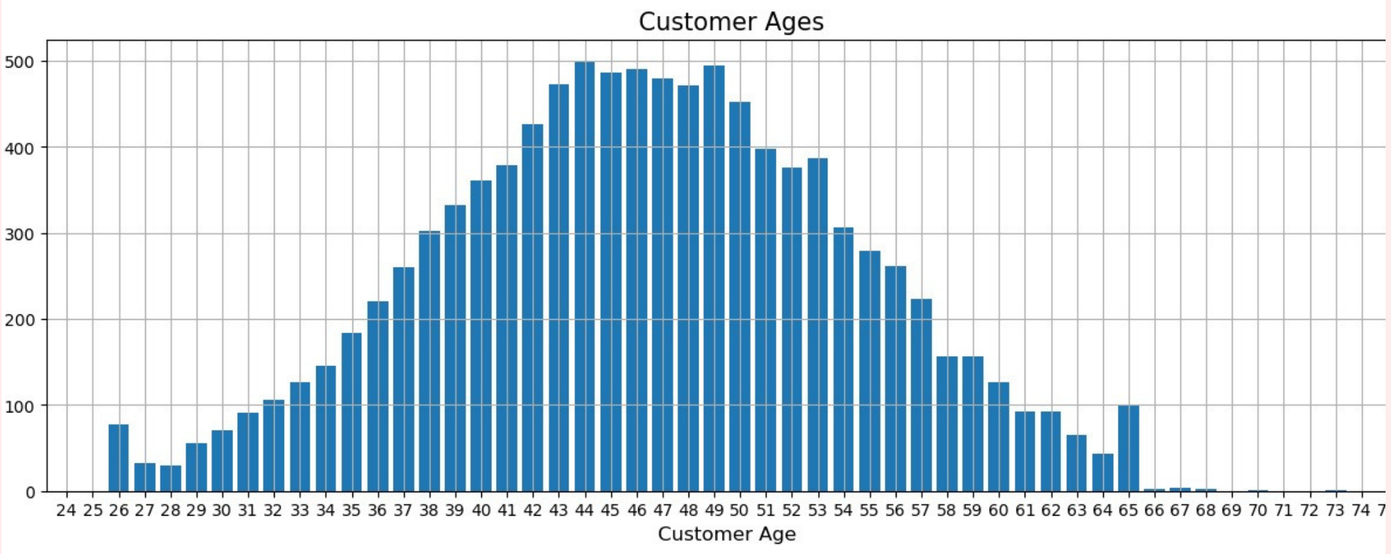
NAIVE BAYES

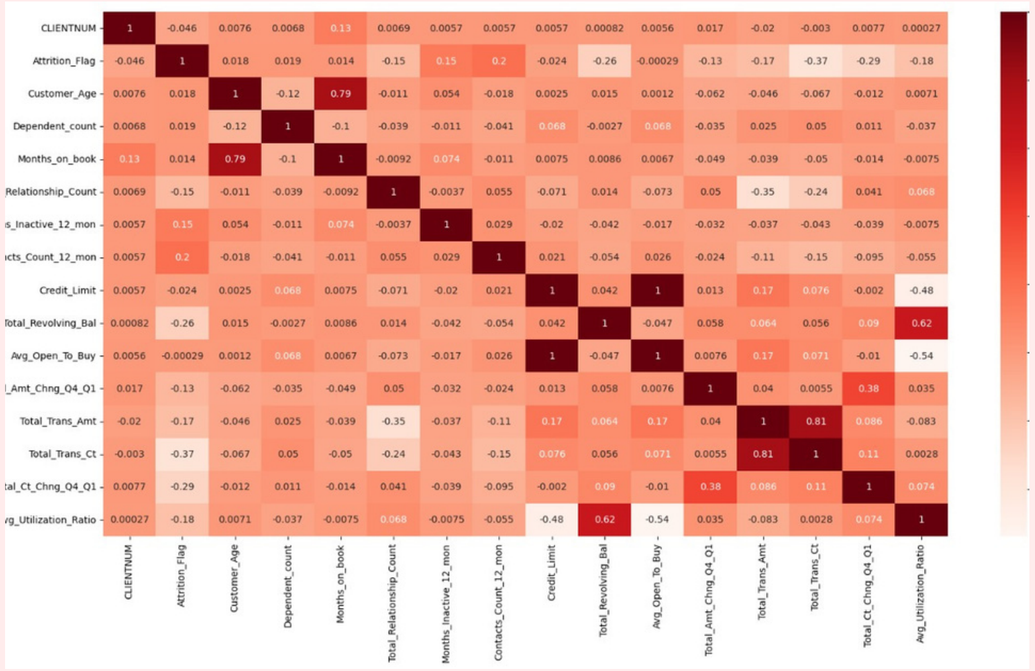
- It is a classification technique based on Bayes' Theorem with an independence assumption among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

- This approach is based on the assumption that the features of the input data are conditionally independent given the class, allowing the algorithm to make predictions quickly and accurately.



RESULTS


Customer Ages
500
400
300
200
100
0
24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 7
Customer Age


CLIENTNUM 1 -0.046 0.0076 0.0068 0.13 0.0069 0.0057 0.0057 0.0057 0.00082 0.0056 0.017 -0.02 -0.003 0.0077 0.00027
Attrition_Flag -0.046 1 0.018 0.019 0.014 -0.15 0.15 0.2 -0.024 -0.26 -0.00029 -0.13 -0.17 -0.37 -0.29 -0.18
Customer_Age 0.0076 0.018 1 -0.12 0.79 -0.011 0.054 -0.018 0.0025 0.015 0.0012 -0.062 -0.046 -0.067 -0.012 0.0071
Dependent_count 0.0068 0.019 -0.12 1 -0.1 -0.039 -0.011 -0.041 0.068 -0.0027 0.068 -0.035 0.025 0.05 0.011 -0.037
Months_on_book 0.13 0.014 0.79 -0.1 1 -0.0092 0.074 -0.011 0.0075 0.0086 0.0067 -0.049 -0.039 -0.05 -0.014 -0.0075
Relationship_Count 0.0069 -0.15 -0.011 -0.039 -0.0092 1 -0.0037 0.055 -0.071 0.014 -0.073 0.05 -0.35 -0.24 0.041 0.068
s_Inactive_12_mon 0.0057 0.15 0.054 -0.011 0.074 -0.0037 1 0.029 -0.02 -0.042 -0.017 -0.032 -0.037 -0.043 -0.039 -0.0075
icts_Count_12_mon 0.0057 0.2 -0.018 -0.041 -0.011 0.055 0.029 1 0.021 -0.054 0.026 -0.024 -0.11 -0.15 -0.095 -0.055
Credit_Limit 0.0057 -0.024 0.0025 0.068 0.0075 -0.071 -0.02 0.021 1 0.042 1 0.013 0.17 0.076 -0.002 -0.48
Total_Revolving_Bal 0.00082 -0.26 0.015 -0.0027 0.0086 0.014 -0.042 -0.054 0.042 1 -0.047 0.058 0.064 0.056 0.09 0.62
Avg_Open_To_Buy 0.0056 -0.00029 0.0012 0.068 0.0067 -0.073 -0.017 0.026 1 -0.047 1 0.0076 0.17 0.071 -0.01 -0.54
l_Amt_Chng_Q4_Q1 0.017 -0.13 -0.062 -0.035 -0.049 0.05 -0.032 -0.024 0.013 0.058 0.0076 1 0.04 0.0055 0.38 0.035
Total_Trans_Amt -0.02 -0.17 -0.046 0.025 -0.039 -0.35 -0.037 -0.11 0.17 0.064 0.17 0.04 1 0.81 0.086 -0.083
Total_Trans_Ct -0.003 -0.37 -0.067 0.05 -0.05 -0.24 -0.043 -0.15 0.076 0.056 0.071 0.0055 0.81 1 0.11 0.0028
tal_Ct_Chng_Q4_Q1 0.0077 -0.29 -0.012 0.011 -0.014 0.041 -0.039 -0.095 -0.002 0.09 -0.01 0.38 0.086 0.11 1 0.074
vg_Utilization_Ratio 0.00027 -0.18 0.0071 -0.037 -0.0075 0.068 -0.0075 -0.055 -0.48 0.62 -0.54 0.035 -0.083 0.0028 0.074 1
CLIENTNUM
Attrition_Flag
Customer_Age
Dependent_count
Months_on_book
Total_Relationship_Count
Months_Inactive_12_mon
Contacts_Count_12_mon
Credit_Limit
Total_Revolving_Bal
Avg_Open_To_Buy
Total_Amt_Chng_Q4_Q1
Total_Trans_Amt
Total_Trans_Ct
Total_Ct_Chng_Q4_Q1
Avg_Utilization_Ratio

RESULTS

```
========== Decision Tree ==========
Accuracy    :  0.8499506416584403
Recall      :  0.8499506416584403
Precision   :  0.8557973340252953
F1 Score    :  0.8499506416584403
Confusion Matrix:
[[2318  250]
 [ 206  265]]
```

```
========== Naive bayes classifier ==========
Accuracy    :  0.8845014807502468
Recall      :  0.8845014807502468
Precision   :  0.8770220536974139
F1 Score    :  0.8845014807502468
Confusion Matrix:
[[2436  132]
 [ 219  252]]
```

CONCLUSION

In conclusion, our focus on the given dataset revolves around distinguishing between existing customers and those who have attrited, essentially left the credit card service.

Our goal is to find patterns and insights in the dataset by utilizing data mining techniques, particularly decision tree analysis and the Naive Bayes algorithm.

Our objective in starting this analytical journey is to project client attrition while also comprehending the underlying variables that influence this choice. With this information, we can better serve our customers, adjust our plans, and make well-informed decisions. Essentially, we are trying to utilize data analytics to differentiate between current and lost clients to promote a more proactive and strategic approach in the credit card industry.


THANK YOU