Credit Card Attrition Prediction

Anitha George, Department of Computer Science
Rajagiri College of Social Sciences, Kochi, Kerala, msccs2207@rajagiri.edu

Abstract— **Credit card attrition refers to the rate at which customers close their credit card accounts. It is a measure of customer churn in the credit card industry. Credit card companies rely on customer retention to maintain profitability, as the cost of acquiring new customers can be high. Therefore, understanding and predicting credit card attrition is crucial for credit card companies to retain their customers and prevent revenue loss. Analyzing customer data and identifying patterns can help credit card companies develop effective retention strategies to reduce attrition rates. This case study explores the problem of credit card attrition and aims to build multiple classification models using Orange tool and select the model with the best accuracy out of it.**

Keywords—credit card, attrition, preprocessing, model, data mining, SVM, kNN, Random Forest, Logistic Regression, Decision Tree.

I. INTRODUCTION

When a consumer cancels their credit card or lets it expire without being renewed, they are said to have attrited from their relationship with the credit card provider. For credit card companies, attrition is a problem because it can result in lost sales and waning client loyalty.

Credit card attrition can be caused by a number of circumstances. Changes in a customer's financial status, dissatisfaction with the credit card's benefits or fees, or the presence of better credit card deals from rival issuers are a few examples of these. Ineffective customer service and ineffective communication from credit card issuers can both lead to attrition.

Credit card issuers may use a number of techniques, such as providing more enticing reward programmes, enhancing customer service, and running targeted marketing efforts to retain consumers, to lower credit card attrition. Exit interviews with customers who have cancelled their cards may also be conducted in order to pinpoint the precise causes of attrition and address these problems moving forward.

Credit card attrition rate is calculated by taking the number of customers lost within a time period and dividing it by the total number of customers at the beginning of this time period and expressed as a percentage of a whole.

II. IMPLEMENTATION

Tool used- Orange.

Orange (3.31.0) is an open-source data visualization and analysis tool. Users can easily perform data preprocessing, data exploration and data analysis using Orange.

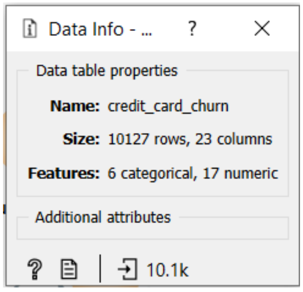
A. Data Description

The dataset used in this project was obtained from Kaggle. It is a sample of credit card customer accounts, along with transactional and demographic data. It contains 10127 instances and 20 features. Out of 21 variables in train data, there are 15 numeric variables and 6 categorical variables. The attributes in the dataset are:

No	*Attribute*	*Description*	*Type*
1	ClientNum	Unique identified for the customer holding the account	Numeric
2	Attrition Flag	Flag indicating if the customer is an attrited or non-attrited customer	Categorical
3	Total Transaction Count	Total number of times the card was used by the account holder	Numeric
4	Total Revolving Balance	Total amount as revolving balance	Numeric
5	Average Utilization Ratio	Average card Utilization ratio	Numeric
6	Total Transaction Amount	Total amount spent by the account holder	Numeric
7	Total Count Change Q4 Q1	Change in transaction count(Q4 over Q1)	Numeric
8	Months Inactive 12 Mon	Number of months inactive in the last 12 months	Numeric
9	Total Relationship Count	Total number of banking products the customer has with the bank	Numeric
10	Contacts Count 12 Month	Number of times the customer contacted the bank in the last 12 months	Numeric
11	Total Amount Change Q4 Q1	Change in transaction amount (Q4 over Q1)	Numeric
12	Average Open To Buy	Open to but credit line (Average of last 12 months)	Numeric
13	Credit Limit	Credit limit of the account holder	Numeric
14	Education Level	Educational qualification of account holder	Categorical
15	Income Category	Annual income category of account holder	Categorical
16	Dependent Count	Number of dependents of account holder	Numeric

B. *Data Preprocessing*

The dataset does not contain any missing value. The target variable chosen here is Attrition_Flag, with two values Attrited customer and Existing customer. It was then converted to binary variable, where the value of Attrited customer was set to 1 and that of the Existing customer was set to 0.



Data Info - ...

Data table properties

Name: credit_card_churn

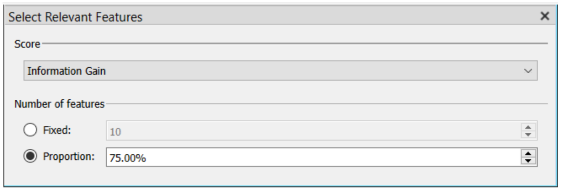
Size: 10127 rows, 23 columns

Features: 6 categorical, 17 numeric

Additional attributes

10.1k

The basic distribution of the data is observed using the feature statistic widget. The attribute ClientNum that does not contribute to the prediction is removed. Rank widget helps to obtain the features that contribute more to the target variable, using information gain, Gini index, etc. The dataset is then preprocessed to remove the features with low value of information gain. Thus preprocessing reduces the number of features to 14 by removing Customer_Age, Dependent_Count, Marital Status, Card_Category and Months_On_Book.



Select Relevant Features

Score

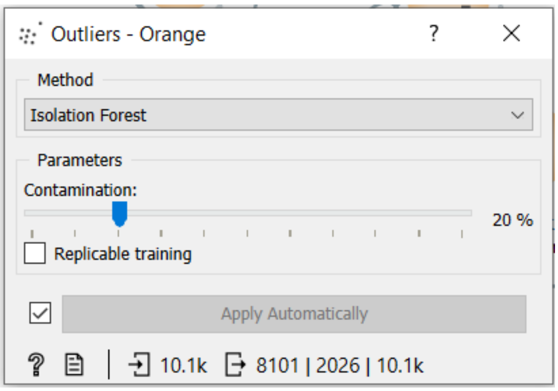
Information Gain

Number of features

○ Fixed: 10

● Proportion: 75.00%

The outliers in the dataset are removed using the outliers widget, where 20% of the contamination (outliers) is removed, resulting in a new dataset of 8101 instances.



Outliers - Orange

Method

Isolation Forest

Parameters

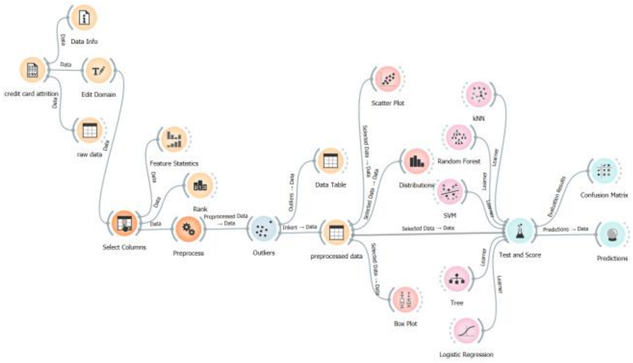
Contamination: 20 %

□ Replicable training

☑ Apply Automatically

10.1k 8101 | 2026 | 10.1k

The following is the Orange connection diagram:



C. *Data Visualization*

Visualization of the dataset helps us to discover patterns, check assumptions and spot weird data points like outliers. The following charts and statistics provide interesting and useful recommendations in prediction.

1)Income category

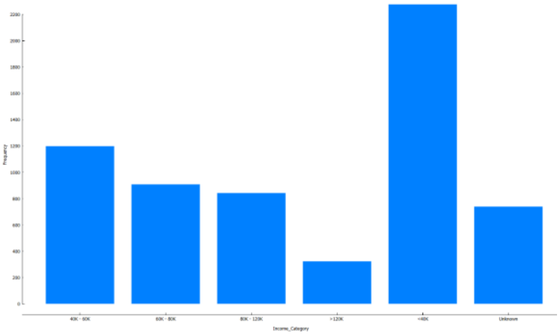


Fig: Income Category

The majority of credit card users were those whose annual income was under 40K. This group is our primary target, therefore rather than concentrating on marketing credit cards to those with high incomes, we stand a higher chance of succeeding by promoting to people with incomes under 40K.

2) Total Transactions

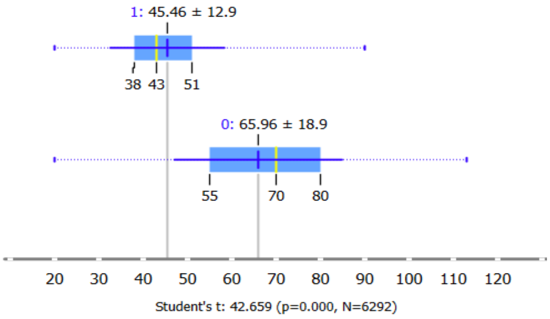


Fig: Total Transactions

Individuals who attrited or churned had had less transactions overall over the last 12 months. We could pay extra attention, such as awarding points for each transaction that can be exchanged for a reward or a voucher, to prevent attrition.

3)Average Credit Utilization vs Credit Limit

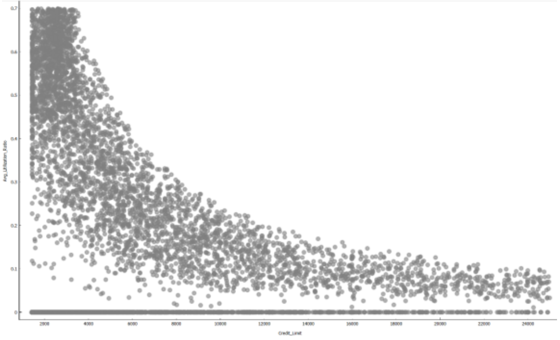


Fig:Plot between average credit utilization and credit limit

As credit limit increases, average card utilization decreases.

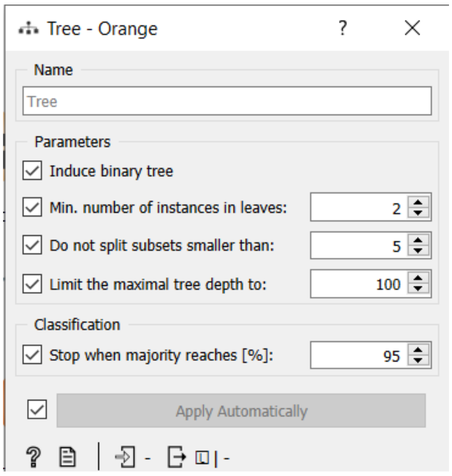
D) Model Creation

Data prediction is done using machine learning algorithms, which consists of a target variable that is to be predicted from a given set of predictors. We create a function that converts input data into the desired outputs using this set of variables. The training procedure is carried out repeatedly until the model's accuracy on the training set reaches the target level.

Logistic Regression, Decision Tree, Random Forest, kNN, and SVM are the algorithms used in this study.

1) Decision Tree

The decision tree algorithm works by recursively partitioning the data into subsets based on the values of the input features, until each subset contains only instances of a single class (in a classification problem) or a single predicted value (in a regression problem). The algorithm chooses the best feature to split on at each node, based on a measure of impurity, such as entropy or Gini index. In a classification problem, the goal is to assign a class label to a given input instance, while in a regression problem, the goal is to predict a continuous numeric value.



Tree - Orange

Name: Tree

Parameters

☑ Induce binary tree

☑ Min. number of instances in leaves: 2

☑ Do not split subsets smaller than: 5

☑ Limit the maximal tree depth to: 100

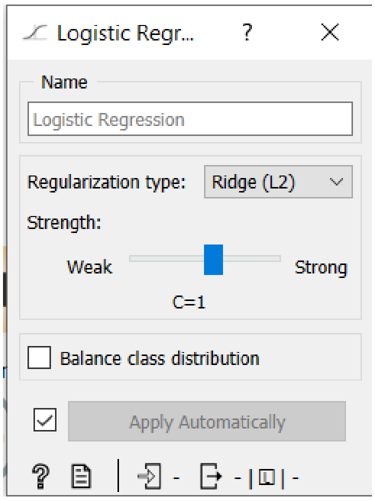
Classification

☑ Stop when majority reaches [%]: 95

☑ Apply Automatically

2) Logistic Regression

It is a classification algorithm used to estimate disrete values based on a given set of independent variable(s). The model uses a logistic function to estimate the probability of the dependent variable. The logistis function produces an S-shaped curve, with values ranging from 0 to 1. The model estimates the coefficients of the independent variables, which are used to calculate the log-odds of the dependent variable. The logit can then be transformed back into probabilities using the logistic function.



Logistic Regr...

Name: Logistic Regression

Regularization type: Ridge (L2)

Strength:

Weak — Strong

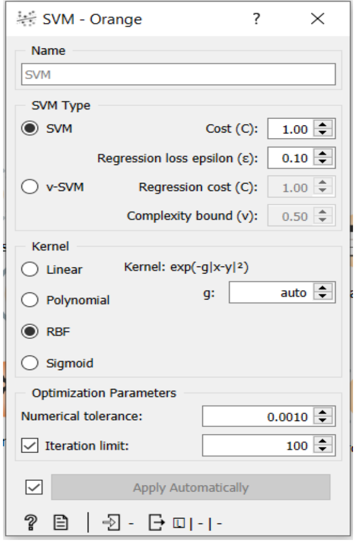
C=1

□ Balance class distribution

☑ Apply Automatically

3) Support Vector Machine (SVM)

The goal of SVM is to find the hyperplane that best separates the data points in a given dataset. In a binary classification problem, SVM algorithm finds the hyperplane that separates the two classes with maximum margin. The margin is the distance between the hyperplane and the closest data points from each class. The hyperplane is chosen such that it maximizes this margin, and it is called the maximum margin hyperplane.



SVM - Orange

Name: SVM

SVM Type

● SVM — Cost (C): 1.00

Regression loss epsilon (ε): 0.10

○ ν-SVM — Regression cost (C): 1.00

Complexity bound (ν): 0.50

Kernel

○ Linear

○ Polynomial

● RBF

○ Sigmoid

Kernel: exp(-g|x-y|²)

g: auto

Optimization Parameters

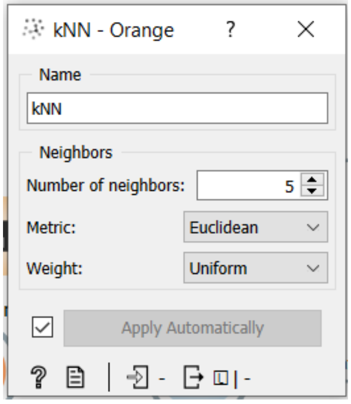
Numerical tolerance: 0.0010

☑ Iteration limit: 100

☑ Apply Automatically

4) *K-Nearest Neighbor*

Both supervised and unsupervised learning can be accomplished with this non-parametric, instance-based learning approach. The k closest training instances in the feature space make up the input for a KNN. A class membership or a forecast value for a new input is the output. KNN locates the k data points that are closest to the new input in order to generate a prediction. It then allocates the new input to the k neighbours' most prevalent class.


kNN - Orange

Name: kNN

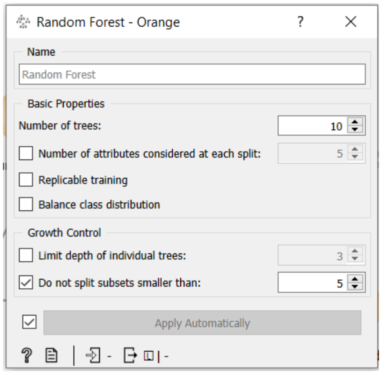
Neighbors
Number of neighbors: 5
Metric: Euclidean
Weight: Uniform

☑ Apply Automatically

5) *Random Forest*

A random selection of the features and various subsets of the training data are used to build a series of decision trees in a random forest. Every decision tree is trained using a portion of the data and characteristics, and the output from every decision tree is then added together to get the final forecast. This ensemble technique aids in lowering overfitting and enhancing the model's generalisation capabilities.

A versatile method, Random Forest may be used to categorical and continuous data and can manage a high number of input features. Additionally, it offers a measure of feature relevance that may be used to determine which properties are most crucial for the model.


Random Forest - Orange

Name: Random Forest

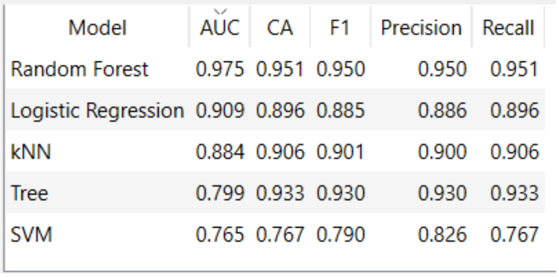
Basic Properties
Number of trees: 10
☐ Number of attributes considered at each split: 5
☐ Replicable training
☐ Balance class distribution

Growth Control
☐ Limit depth of individual trees: 3
☑ Do not split subsets smaller than: 5

☑ Apply Automatically

III. RESULT

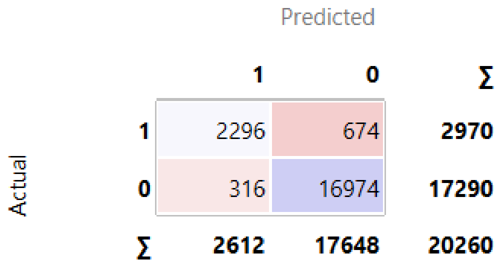
Model evaluation value of each model is shown below:



Model	AUC	CA	F1	Precision	Recall
Random Forest	0.975	0.951	0.950	0.950	0.951
Logistic Regression	0.909	0.896	0.885	0.886	0.896
kNN	0.884	0.906	0.901	0.900	0.906
Tree	0.799	0.933	0.930	0.930	0.933
SVM	0.765	0.767	0.790	0.826	0.767

It is evident from the data that these models can be used to predict credit card attrition with high probability.

Among the applied classification models, Random Forest has the highest accuracy. The confusion matrix of Random Forest is shown below:


Predicted

Actual

	1	0	∑
1	2296	674	2970
0	316	16974	17290
∑	2612	17648	20260

IV. CONCLUSION

The machine learning model created using Random Forest algorithm can predict attrition with an accuracy of 95.11%. The number of transactions in the last 12 months is considered to be the most influential feature to predict attrition. People with less than 40K income is the potential customer, and this category should be prioritized while marketing. Through active engagement and campaigns, the company should concentrate on retaining its top customers.

REFERENCES

[1] Most Common Machine Learning Algorithms With Python & R Code (analyticsvidhya.com)

[2] https://blog.devgenius.io/credit-card-customer-churn-predictive-analytics-b012ff8c385d

[3] https://vertical-institute-assets.s3.ap-southeast-1.amazonaws.com/Data+Analytics+Bootcamp+Capstone+Project+by+Lee+Ying+Chia.pdf

[4]] AL-Najjar, D.; Al-Rousan, N.; AL-Najjar, H. Machine Learning to Develop Credit Card Customer Churn Prediction. J. Theor. Appl. Electron. Commer. Res. 2022, 17, 1529–1542

[5] Ünlü, Kamil Demirberk. (2021). Predicting credit card customer churn using support vector machine based on Bayesian optimization. Communications Faculty Of Science University of Ankara. 70. 827-836. 10.31801/cfsuasmas.899206.