CSE 587 Data Intensive Computing

Name : Govardhan Arikatla
Person number : 50449747
Ubit: garikatl

Analysis of credit cards attrition data in predicting the status of customer attrition status

Problem Statement
Customer attrition can be due to various reasons that impact customers considering their livelihood and background. In this project, with the help of the credit card attrition dataset, the factors influencing customer attrition are identified and analyzed. This information can be helpful for credit card companies to know the reasons behind their customer's decline and implement programs to improve the customer experience and reduce customer attrition. This project deals with the prediction of customer attrition from credit companies based on the given factors of individual customers.
Dataset link:
https://www.kaggle.com/code/santij95/predicting-credit-cards-customer-attrition

Modeling:

To fit the machine learning models, the effective categorical columns in the dataset have been labeled using a labeling encoder and all the columns have been scaled and transformed into the same range.
The imbalances of the data in the dataset for attrited customers have been sampled using SMOTE technique. In general SMOTE creates the new examples for the minority class in the data.

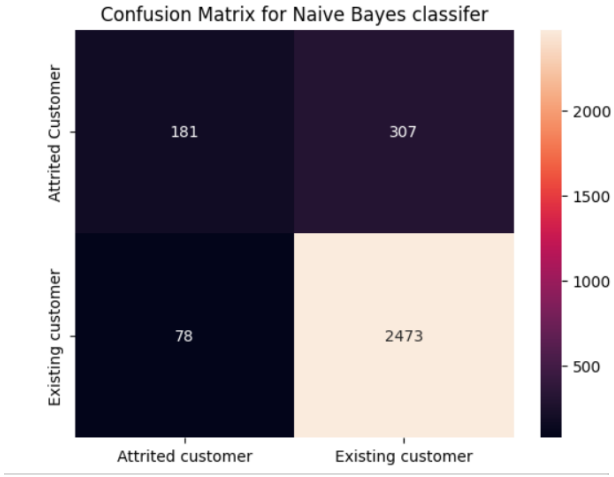
The machine learning algorithms used in this project are:

1. Naive Bayes algorithm
2. Logistic Regression algorithm
3. KNN classifier algorithm'
4. Decision tree classifier algorithm
5. XGBoost algorithm

1. Naive Bayes algorithm
Why naive bayes on credit card attrition data?

Naive Bayes algorithm can be a good fit if there are discrete or categorical independent features in the data. With our data being the independent features of Income category and dependent count as categorical, naive bayes algorithm can be used to fit the model with credit cards attrition dataset and helps to classify the customer as attrited or existing customers.

The model developed with naive bayes has been used for the prediction of the test dataset and the confusion matrix is visualized.

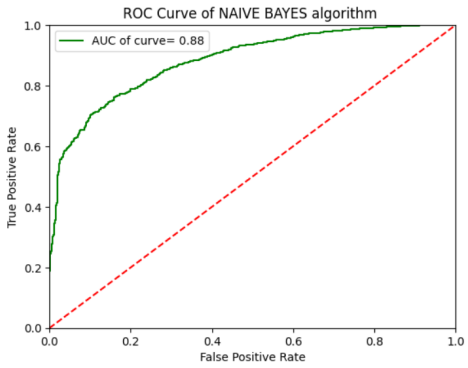


The statistics of the naive bayes algorithm model is shown in the figure. Accuracy, precision, recall and f1 score are considered for evaluating the model.

```
from sklearn.metrics import classification_report
print(classification_report(y_test, gnb.predict(X_test)))
```

	precision	recall	f1-score	support
Attrited Customer	0.70	0.37	0.48	488
Existing Customer	0.89	0.97	0.93	2551
accuracy			0.87	3039
macro avg	0.79	0.67	0.71	3039
weighted avg	0.86	0.87	0.86	3039

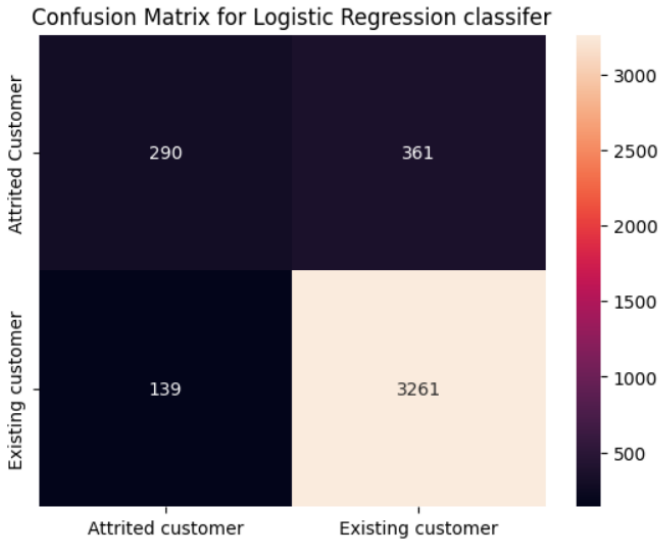
The true positive and true negative values are plotted using the ROC curve.



2. Logistic Regression algorithm

Why is the logistic regression algorithm applied on credit card attrition data??
Logistic regression by nature is well suited for the binary classification problems provided with the independent input variables. As our data deals with binary outcome of predicting whether the customer is attrited or not from the credit card companies based on the independent features average utilization ratio,Dependent count, Contacts count in 12 months etc, So, the Logistic Regression algorithm is developed using the credit cards attrition data and the results of the model are analyzed.

The confusion matrix for the prediction if attrited customer is shown as below.

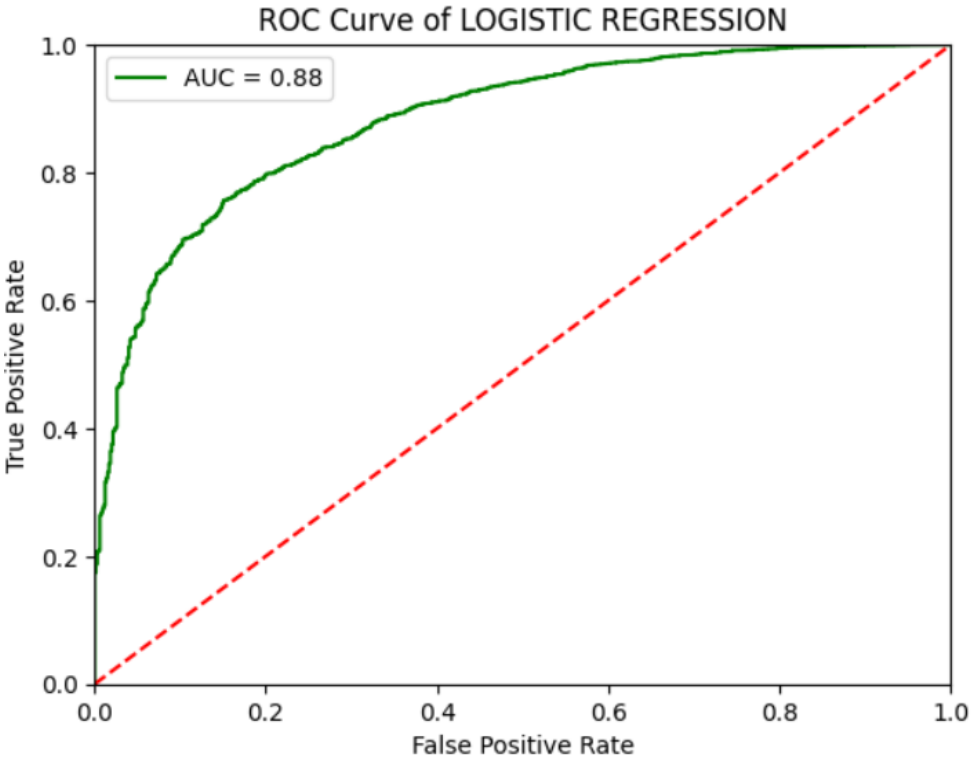


The accuracy, precision, recall, and f1 score values of the logistic regression model are:

```
print(classification_report(y_test2, lr.predict(X_test2)))
```

	precision	recall	f1-score	support
Attrited Customer	0.68	0.45	0.54	651
Existing Customer	0.90	0.96	0.93	3400
accuracy			0.88	4051
macro avg	0.79	0.70	0.73	4051
weighted avg	0.86	0.88	0.87	4051

The visualization of the ROC curve is shown in the below figure.

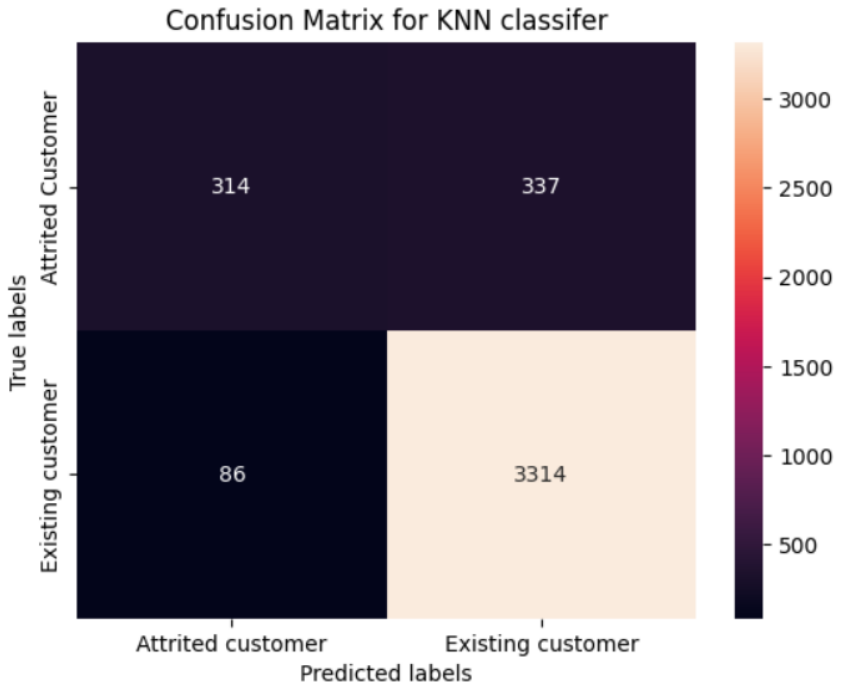


3. K Nearest Neighbors Classifier algorithm

Why KNN algorithm on Credit card attrition data?

As the K Nearest Neighbors classifier deals better with large numbers of independent variables, and with less complex relationships between independent variables, our data comes under the same category. So, KNN algorithm can be applied to develop the model with credit cards attrition data and predict the status of attrition with given independent variables as input.

The visualization of confusion matrix results are:

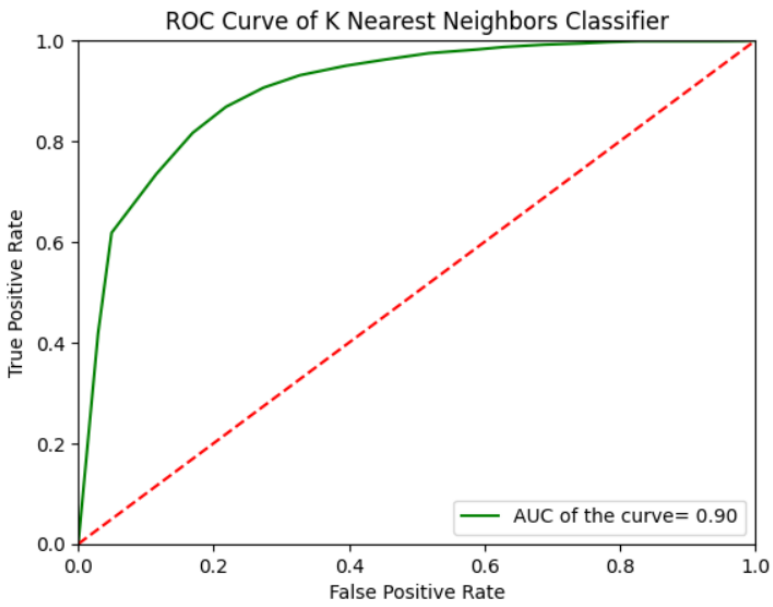


The evaluation metrics of the model are:

```
print(classification_report(y_test3, knn.predict(X_test3)))
```

	precision	recall	f1-score	support
Attrited Customer	0.79	0.48	0.60	651
Existing Customer	0.91	0.97	0.94	3400
accuracy			0.90	4051
macro avg	0.85	0.73	0.77	4051
weighted avg	0.89	0.90	0.88	4051

The visualization of the ROC curve is



4. Decision Tree Classifier algorithm

Decision Tree classifier is one of the supervised machine learning algorithms that can be used to make predictions in both classification and regression cases. Nodes and leaves make up a decision tree, where nodes represent the condition on the characteristics in a dataset and leaves represent the result of that condition. To split or categorize the attributes in a dataset, decision trees use a top-down method. They use several criteria to define the splits or classes. Entropy, information gain, and Gini index are a few of them. Decision tree algorithms split the dataset into different datasets based on the independent variables.

Why the decision tree algorithm on credit card attrition data?
The Decision tree algorithm is implemented on the Credit card customer attrition dataset to predict the status of the attrition of the customers from the credit card companies. It takes the features such as average utilization of the credit, total dependents of the customer from the dataset that are mostly affecting the customer attrition for determining the output of attrition.

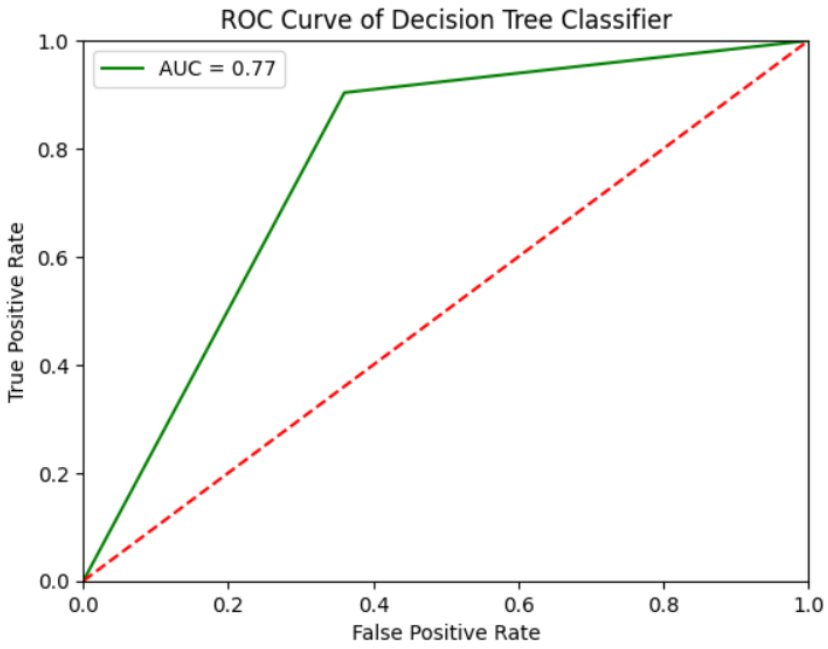
The confusion matrix of true and predicted values of Attrited and existing customers using decision trees is plotted as below:

The metrics Accuracy, F1 score, Precision and Recall for the decision tree classifier are calculated. The values of the resulting metrics are shown below.

```
print(classification_report(y_test4, y_pred4))
```

	precision	recall	f1-score	support
Attrited Customer	0.56	0.64	0.60	325
Existing Customer	0.93	0.90	0.92	1701
accuracy			0.86	2026
macro avg	0.74	0.77	0.76	2026
weighted avg	0.87	0.86	0.87	2026

The plot of ROC curve is as below:

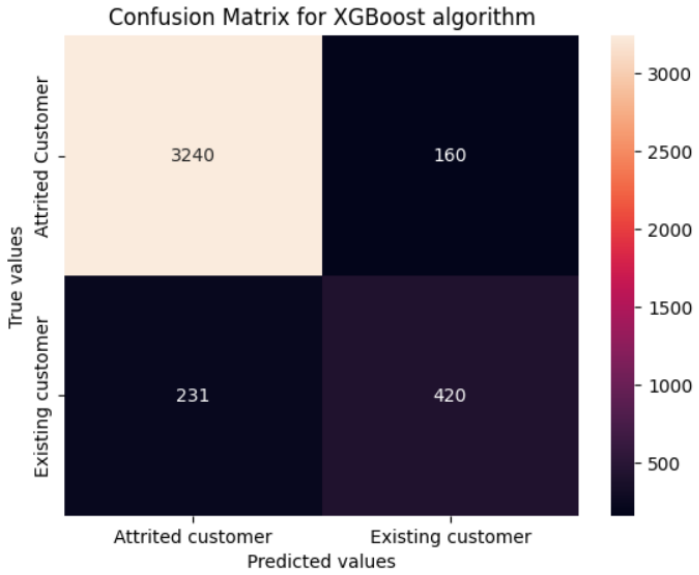


5. XGBoost algorithm

Extreme Gradient Boost or XGBoost is an extended version of Gradient Boosting algorithm which can be applied for solving both classification and regression problems. XGBoost is a supervised machine learning algorithm with input and labels. It uses the method of ensemble learning for training the model. XGBoost algorithm aggregates the results of multiple week learning models and creates an optimized machine learning model for classification or regression. The XGBoost algorithm is well known for handling missing values in the data and also in dealing with larger datasets.

Why the XGBoost algorithm on credit cards attrition data?
The XGBoost algorithm works better if there are imbalances in the data. In our data, the number of attrited customers is less than the number of existing customers, So XGBoost algorithm is used for developing the classification model using the independent variables such as Dependent count, Total relationship count, Months inactive in last 12 months etc. and predicting whether the customers are most probably attrited or existing.

The confusion matrix of true and predicted values of Attrited and existing customers is visualized as below:

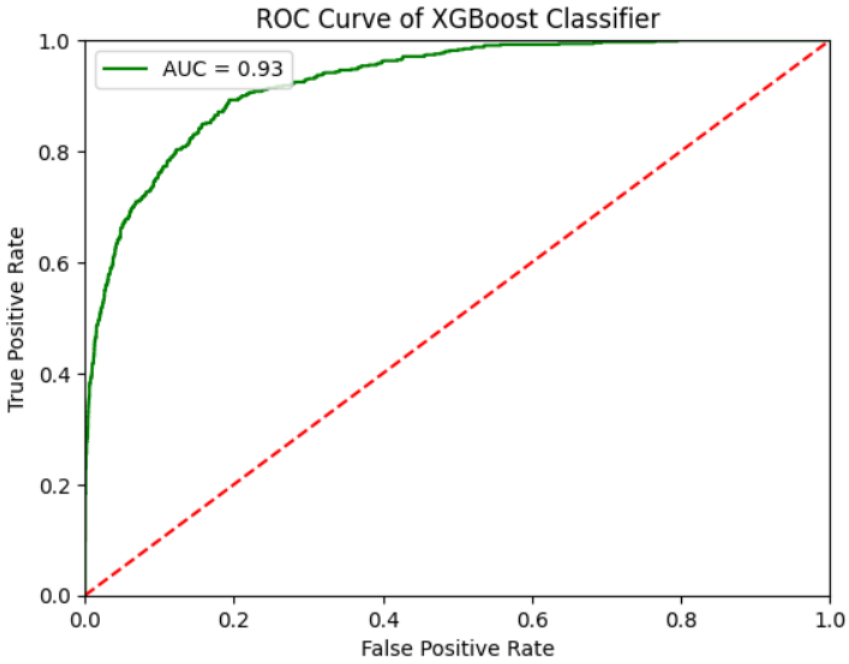


The model developed using XGBoost algorithm has been evaluated using the metrics Accuracy, F1 score, Precision and Recall. The values of the resulting metrics are shown below.

```
print(classification_report(y_test5,y_pred5))
```

	precision	recall	f1-score	support
0	0.93	0.95	0.94	3400
1	0.72	0.65	0.68	651
accuracy			0.90	4051
macro avg	0.83	0.80	0.81	4051
weighted avg	0.90	0.90	0.90	4051

The ROC curve for the XGboost algorithm is shown in the below figure.



References

1. https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d
2. https://towardsdatascience.com/decision-tree-classifier-explained-in-real-life-picking-a-vacation-destination-6226b2b60575
3. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html
4. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_curve.html
5. https://stackoverflow.com/questions/25009284/how-to-plot-roc-curve-in-python