Retention of credit card users

Kamal Afridi M Shaik, Sreenivas Lebaka
myduku98@students.rowan.edu
lebaka98@students.rowan.edu
Rowan University
Glassboro, NJ, USA

ABSTRACT

Customer retention is a vital problem in every company. Every client retention is an essential indicator that reveals the reality about overall customer retention rates. A thorough examination of existing models for forecasting customer retention is conducted in this project, and four models based on Machine Learning algorithms is proposed for predicting customer turnover in the banking industry. The suggested machine learning process combines methods such as kNN, Decision Tree, SVM, and Random Forest. The performance measures that were examined between these models built were accuracy, precision, recall, and F1 score. The Random-Forest Classifier model has been found to perform better than other models.

KEYWORDS

Customer Retention, kNN, Decision Tree, Random forest, SVM, Prediction

Reference Format:
Kamal Afridi M Shaik, Sreenivas Lebaka. 2022. Retention of credit card users. In *Archive '22*. USA, 6 pages.

1 INTRODUCTION

Customers are the foundation of every company. Retaining current customers is a big issue in today's competitive market. Various banking companies have entered the industry, making it essential for banks to enhance the quality of services they provide to retain users.[1] Almost every service provider has introduced modern technologies that help them to provide clients with simple ways to save money or do transactions. The customer retention process is handled by the company's customer relationship management department consisting of data scientists and analysts. As a result, the executive officers' choice to adjust the operations or expand their services in order to keep current consumers or attract new ones is important.

(1) Customers look for a variety of features in a bank service provider, such as geographical proximity, online services, transaction feasibility, account flexibility, rates, and policies, among others. In order to keep customers, banks must track their competitors' initiatives to enhance them to use their services. As a result, they must make decisions in order to stay in business; else, companies might lose customers soon.
(2) Banks employ analysts to create reports based on client data in order to keep track of their customers. Data on both current and defaulting customers will often include transaction details, services customers are actively utilizing, personal information such as family details, employment, and income, and credit records. Data analysts deal with the data by applying algorithms, creating predictions, and generating reports that show the bank's customers' operations. These insights are vital in making decisions for effective retention-related outcomes.
(3) This project analyzes the data and predicts whether a client will retain or not based on a variety of parameters.

Archive 2022, May 06, 2022, .

2 RELATED WORKS

Customer retention prediction approaches may be divided into two categories: prediction methods based on classical statistical analysis,and prediction methods based on machine learning. Linear discriminant analysis, the naïve Bayesian model, cluster analysis, and logistic regression are some of the commonly used statistical analysis-based prediction approaches. Pinar et al., for example, utilized a naïve Bayes classifier in 2011 to forecast customer turnover for a telecom business. Customers' average call time was substantially connected with customer retention, according to their study[2]. Renjith et al. used logistic regression to predict e-commerce customer attrition and offered a customised client retention approach based on machine learning [3]. However, these methods have a limitation while dealing with big data and multidimensional variable data; the prediction performance was not obvious in these cases.

Prediction methods based on machine learning mainly include decision trees, the support vector machine, and artificial neural networks, among others. Decision trees, for example, are commonly employed in practical customer retention prediction, according to Neslin et al. The decision tree technique may be used to predict loss as a simple model [4]. Zhang et al. applied the C5.0 decision tree to anticipate the loss of telecom businesses' postal short messaging service. The C5.0 decision tree prediction model was shown to have a good level of accuracy [5].

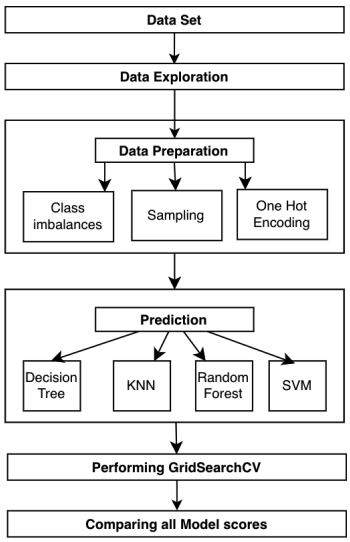
3 DATA SET

The Project is performed by utilizing a dataset of bank customer data. The dataset contains information on 10128 customers with 21 attributes with no null values in any record, they are divided into two groups: those who exit (represented by 1 in the target variable) and those who did not leave (represented by 0 in the target variable). The dataset includes 1627 clients who left the bank services, with the remainder remaining. Individual client data include annual income, marital status, gender, age, utilization period, opening

balance, number of dependents, values indicating if the customer has a credit card and card category, values indicating whether the customer is an active member, and finally the customer's credit balance.

```
Data columns (total 21 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   CLIENTNUM                 10127 non-null  int64
 1   Attrition_Flag            10127 non-null  category
 2   Customer_Age              10127 non-null  category
 3   Gender                    10127 non-null  category
 4   Dependent_count           10127 non-null  int64
 5   Education_Level           10127 non-null  category
 6   Marital_Status            10127 non-null  category
 7   Income_Category           10127 non-null  category
 8   Card_Category             10127 non-null  category
 9   Months_on_book            10127 non-null  int64
 10  Total_Relationship_Count  10127 non-null  int64
 11  Months_Inactive_12_mon    10127 non-null  int64
 12  Contacts_Count_12_mon     10127 non-null  int64
 13  Credit_Limit              10127 non-null  float64
 14  Total_Revolving_Bal       10127 non-null  int64
 15  Avg_Open_To_Buy           10127 non-null  float64
 16  Total_Amt_Chng_Q4_Q1      10127 non-null  float64
 17  Total_Trans_Amt           10127 non-null  int64
 18  Total_Trans_Ct            10127 non-null  int64
 19  Total_Ct_Chng_Q4_Q1       10127 non-null  float64
 20  Avg_Utilization_Ratio     10127 non-null  float64
dtypes: category(7), float64(5), int64(9)
```

4 METHODOLOGY



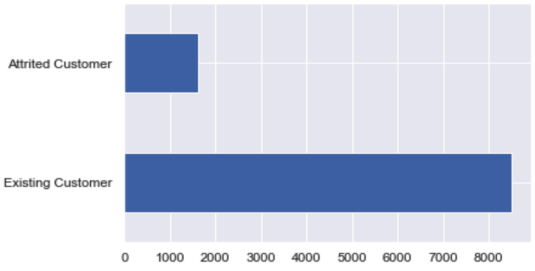
(1) **Data exploration:**
To gain a better understanding of the data, numerous analyses were run on it using Python in Jupyter Notebook IDE. In the example of the 'Customer Age' variable, four bins were constructed and assigned the values to the appropriate class. The remaining category variables are listed below.

```
Customer_Age: ['0_35', '35_45', '45_55', '55_70']
Gender: ['F', 'M']
Education_Level: ['College', 'Doctorate', 'Graduate', 'High School', 'Post-Graduate', 'Uneducated', 'Unknown']
Marital_Status: ['Divorced', 'Married', 'Single']
Income_Category: ['$120K +', '$40K - $60K', '$60K - $80K', '$80K - $120K', 'Less than $40K', 'Unknown']
Card_Category: ['Blue', 'Gold', 'Platinum', 'Silver']
```

(2) **Data preparation:**

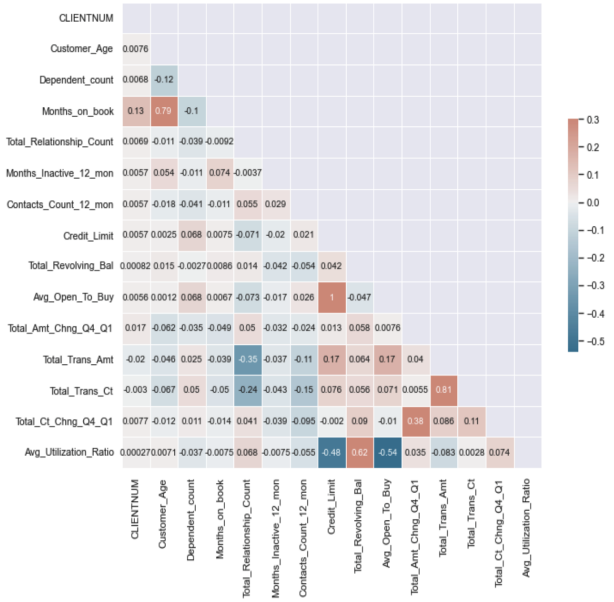
- *Class imbalance problem:*

To cope with the problem of class imbalance, Stratified K-Fold approach was employed. The class 'Existing Customer' or label '0' is far more widespread than the class 'Attired Customer' or label '1', as seen in the graph below. As a result, we can ensure that the models are trained for both class labels by stratifying the target variable.



- *Correlation matrix:*

The correlation matrix gives the correlation coefficient between every two attributes in the data. Using seaborn library's heat map, we can plot the correlation matrix.



- *Sampling technique:*

In this project, K-Fold cross validation strategy was used as the sampling technique. In K-Fold, the original dataset is shuffled and randomly partitioned into k equal-sized subsets called folds. A single subset from those k- subset is kept aside for validation of the testing the model, while the remaining (k-1) subsets are utilized to train the model.

In this project 10-folds or subsets were created through which the models are rigorously trained. To evaluate the model performance cross validation score is used by plotting the confusion matrix for the models.

- *One-hot encoding:*

For the categorical variables in the dataset, one-hot encoding yields a sparse matrix. It produces a new column for each and every record for every category in the dataset, and then merges it with the original data set. To create the final dataset for model training, the category columns are removed. Pandas was used for get-dummies function to create the sparse matrix in this project.

5 EXPERIMENT AND RESULTS

Four models were applied to fit the data in this research. GridSearchCV from Sklearn was used to discover the optimum parameters, model, and scores. GridSearchCV makes use of a user-defined set of hyperparameters. The model is rigorously trained using the training data using these parameter combinations

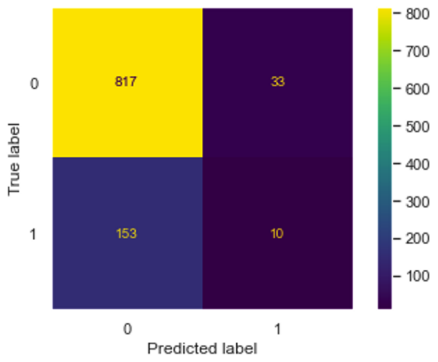
- *K-nearest neighbor classifier*

In this model, classification is carried out by a simple majority voting strategy of each point's nearest neighbors: a query point is allocated to the data class having the most representation among the point's nearest neighbors. Below image is the best estimator for k-Nearest Neighbors according to GridSearchCV.

```
KNeighborsClassifier(n_neighbors=100, p=1)
```

The classification report for this model is shown below:

```
Classification report:
              precision    recall  f1-score   support

           0       0.84      0.96      0.90       850
           1       0.23      0.06      0.10       163

    accuracy                           0.82      1013
   macro avg       0.54      0.51      0.50      1013
weighted avg       0.74      0.82      0.77      1013
```

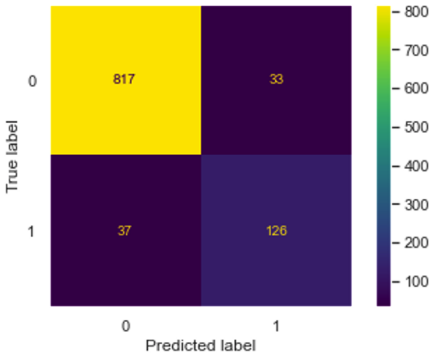


- *Decision tree*

Decision tree is the most common yet best model for classification tasks. GridSearchCV utilizes decision tree parameters such as tree depth, max leaf nodes, random-states to find the best parameters for the model.
The classification report for this model is shown below:

```
DecisionTreeClassifier(max_depth=5, max_leaf_nodes=16, random_state=0)
```

```
Classification report:
              precision    recall  f1-score   support

           0       0.96      0.96      0.96       850
           1       0.79      0.77      0.78       163

    accuracy                           0.93      1013
   macro avg       0.87      0.87      0.87      1013
weighted avg       0.93      0.93      0.93      1013
```



- *Support Vector Machine*

By showing the hyperplane in the 2D plot of input and output variables, the support vector machine employs kernels to predict output class labels. The GridSearchCV may be used to define the model's parameters.

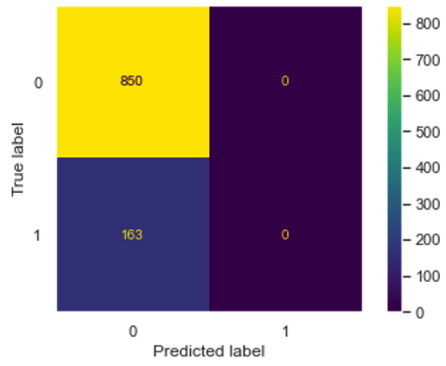
```
SVC(C=0.1, gamma=1)
```

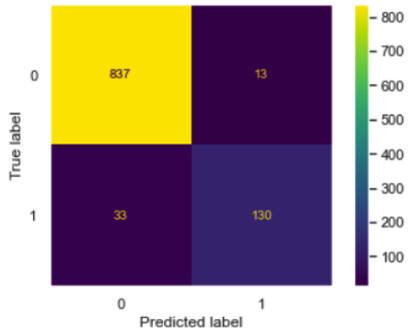
```
print(grid.best_params_)
```

```
{'C': 0.1, 'gamma': 1, 'kernel': 'rbf'}
```

The classification report for this model is shown below:

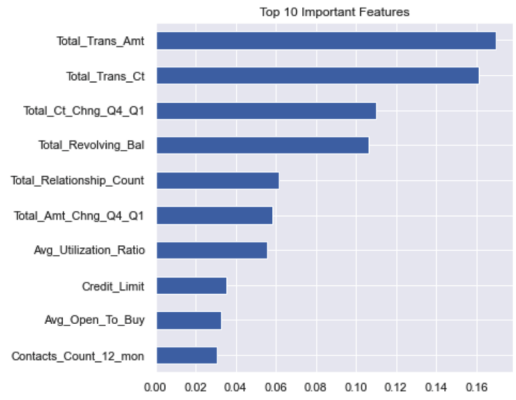
```
Classification report for SVC:
              precision    recall  f1-score   support

           0       0.84      1.00      0.91       850
           1       0.00      0.00      0.00       163

    accuracy                           0.84      1013
   macro avg       0.42      0.50      0.46      1013
weighted avg       0.70      0.84      0.77      1013
```





- *Random-Forest Classifier*

To train and evaluate the data, Random Forest employs numerous decision trees. We may draw the attribute significance graph for predicting the output variable using this model.



The classification report for this model is shown below:

```
RFC performance report:
              precision    recall  f1-score   support

           0       0.96      0.98      0.97       850
           1       0.91      0.80      0.85       163

    accuracy                           0.95      1013
   macro avg       0.94      0.89      0.91      1013
weighted avg       0.95      0.95      0.95      1013
```

6 CONCLUSION

Comparing Model Performances:

The model performances were evaluated with Accuracies, Cross Validation scores, F1 scores and Precision and compared using bar plots.

The results obtained from the different models and the comparisons of the results are provided in the next section

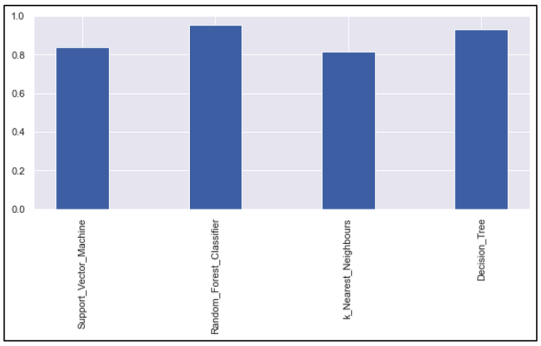


Fig 1. Comparing **Accuracies**

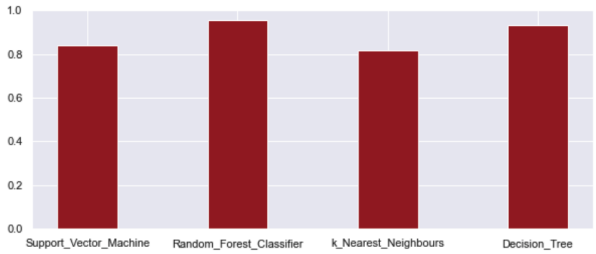


Fig 2. Comparing **Cross validation scores**

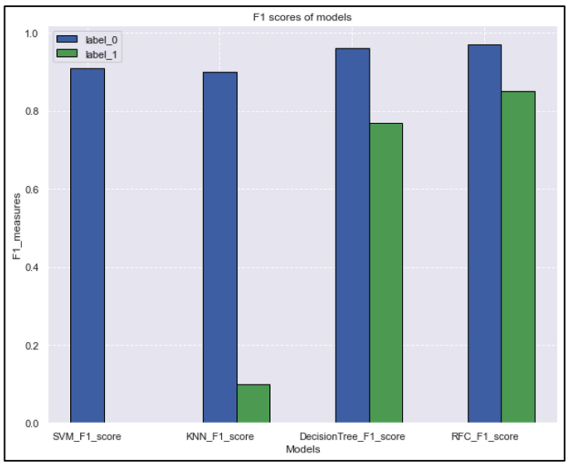


Fig 3. Comparing **F1 Scores**

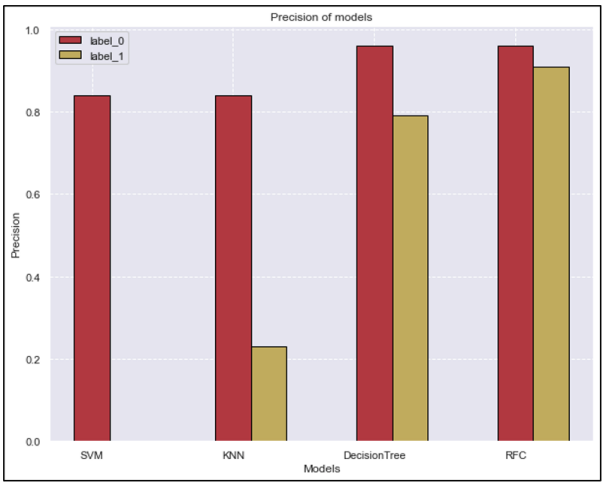


Fig 4. Comparing **Precisions**

We can observe in the above plots that Random-Forest Classifier was the best model since it performed the best in all scoring measures.

REFERENCES

[1]"Bank Customer Retention Prediction And Customer Ranking Based On Deep Neural Networks", International Journal of Science Engineering Development Research (www.ijsdr.org), ISSN:2455-2631, Vol.5, Issue 9, page no.444 - 449, September-2020

[2] Pınar, K.; Topcu, Y.I. Applying Bayesian Belief Network approach to customer chum analysis: A case study on the telecom industry of Turkey. Expert. Syst. Appl. 2011, 38, 7151–7157. Intl. J. Eng. Trends Technol. 2015, 27, 152–157.

[3]Renjith, S. An integrated framework to recommend personalized retention actions to control B2C E-commerce customer churn.

[4]Neslin, S.A.; Gupta, S.; Kamakura, W.; Lu, J.; Mason, C.H. Defection detection: Measuring and understanding the predictive accuracy of customer churn models. J. Mark. Res. 2006, 43, 204–211.

[5]Zhang, Y. A Customer Churn Alarm Model based on the C5.0 Decision Tree-Taking the Postal Short Message as an Example. Stat. Inf. Forum. 2015, 30, 89–94.

7 ACKNOWLEDGEMENT

I cannot express enough thanks to my professor for his continued support and encouragement: Mr. Ashish Chanda.

My completion of this project could not have been accomplished without the support of my project partner, Sreenivas Lebaka.

8 CONTRIBUTION

We as a team worked on this project together, sharing ideas back and forth for 4 weeks.

Data exploration, data processing, Visualization, Sampling, One hot encoding is done by Sreenivas.

Model building, Model evaluation, Parameter tuning, GridSearch CV, Scores comaprision is done by Kamal.