Customer Churn Prediction using Classification Machine Learning Models

Aarushee Thakur
MBA-Business Analytics and Big Data
LM Thapar School of Management
Derabassi - Barwala Rd, Chandigarh, Punjab 140507
Email: athakur$_m ba22@thapar.edu$

Rimjhim Sharma
MBA-Business Analytics and Big Data
LM Thapar School of Management
Derabassi - Barwala Rd, Chandigarh, Punjab 140507
Email: rsharma1$_m ba22@thapar.edu$

***Abstract*—Customer churn is a major problem and one of the most important concerns for large companies. Due to the direct effect on the revenues of the companies, especially in the Banking field, banks are seeking to develop means to predict potential customer to churn. Therefore, finding factors that increase customer churn is important to take necessary actions to reduce this churn. The main contribution of our work is to develop a churn prediction model using classifications models of machine learning like Logistic Regression, KNN, Decision Tree which assists bank operators to predict customers who are most likely subject to churn.**

I. Introduction

The people around especially the customers are surrounded by alot of information and to gain those information large sources are avaliable. Smartphones,e.g., provide instant access to various branded products, mbanking, comparative information. And customer perspectives demand based on advanced technology,convenience reasons, price sensitivity, service quality keeps on changing. Having so many choices, the customers tend to shift from one service to another.

For banking sector, this customer churn prediction is the serious issue and gargantuan impact on the profit line of bankers. Thus, customer retention scheme can be targeted on high-risk customers who wish to discontinue their custom and switch to another competitor.

The paper presents some Classification supervised machine learning models to classify the bank customers into two categories/classes, whether they will leave or not.

II. Related Work

From the above discussion, it is clear that customer retention is important for acompany and for its business strategy. Customer churning becomes business intelligence to know which customers will shift or who will get retained. To achieve customer churning, companies started adapting machine learning techniques for customer churn prediction models. In this section, a few techniques are compared considering churn prediction.

Yaya Xie,XiuLi,E.W.T.Ngai,WeiyunYing proposed a IBRF (Improved Balances Random Forest) methodology to predict the customer churn of the banking sector.Using this model, the reason ofcustomer leaving the bank can be easily acquainted by entering the parameters.

'Hend Sayed' presented a methodology of decision tree in which two packagesML and MLib were conducted, to evaluate accuracy, model training and modelevaluation. They got effective result with ML package.

Hajar Ghaneej and Sayed Mohammad Mirmohammadi applied cross industry data mining techniques Random sampled data of 4383 customers of electronic banking services from the bank's database.

III. Dataset

Dataset used for this project is acquired from an online source, Kaggle. The dataset is subjected to churning of customers of bank containing information about 10,000 customers with 14 features for each customer. The customers of the bank are identified as churn or loyal based on the potential features like credit score,age, gender, estimated salary, etc. A user of the bank is classified as loyal if he/sheis active and remains with the bank. Customers are classified as churners if they switch to another bank. The variable exited in the dataset gives the actual status of the customer if he/she had switched to another bank.

IV. Data Preprocessing

A. Libraries

1) Numpy: NumPy is the fundamental package for scientific computing in Python. NumPy arrays facilitate advanced mathematical and other types of operations on large numbers of data.

2) Pandas: Pandas is an open source Python package that is most widely used for data science/data analysis and machine learning tasks. It is built on top of another package named Numpy, which provides support for multi-dimensional arrays.

3) Matplotlib and Seaborn: Matplotlib and Seaborn both are data visualiztion libraries is used for ploting the data.

4) Sklearn: Provides a selection of efficient tools for machine learning and statistical modeling.

The dataset had no missing values and the categorical data was handles using one hot encoding. Using train test split from sklearn library and the data was split into 25 perecnt testing and the remaining on training.

B. Feature Scalingg

Feature Scalimg is used to standardize the independent variable by keeping the variables range same such that one varibale does not dominate the other.

Fig. 1. Data Preprocessing

V. Data Visualization

Using Seaborn and Matplotlib the following visualizations helps for a better understanding of the data features and their dependency on each other.

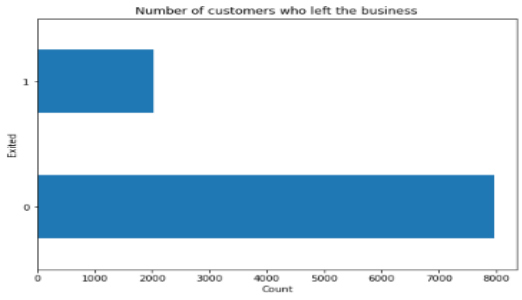

Fig. 2. Number of customers who left the business

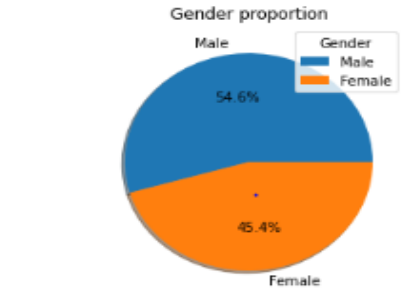

Fig. 3. Gender proportion

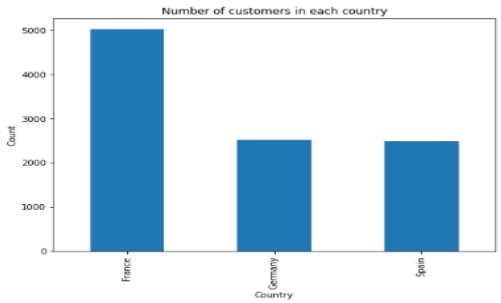

Fig. 4. Total customers in each country

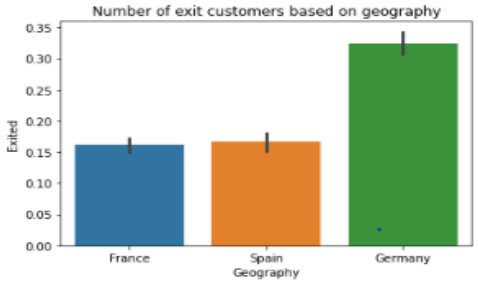

Fig. 5. Number of exit customers from each geography

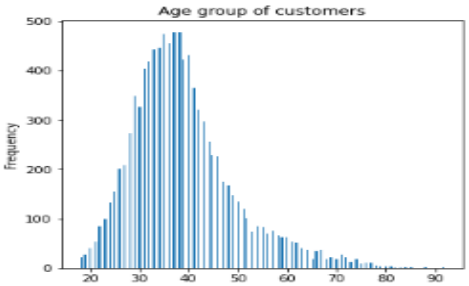

Fig. 6. Age distribution

VI. Methodology

The focus of this project is to predict whether the customer has left the bank or not and by using Supervised Classification Machine Learning algorithms.

A. Logistic Regression

Logistic regression, where the target is categorical, is another effective supervised machine learning method for binary classification problems.A S shaped logistic function is used for fitting the data points.The probabilistic values of the model lie between 0 and 1.

$$\text{Logistic Function} = \frac{1}{1+e^{-y}}$$

Fig. 7. Logistic (sigmoid) Function

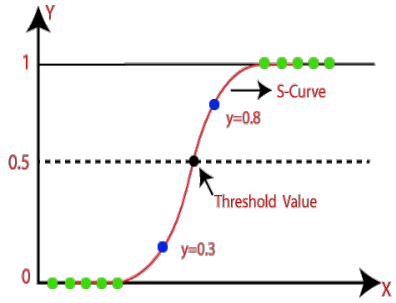

Fig. 8. Logistic (sigmoid) Graph

Using a confusion matrix the performance of the model can be determined.

- Precision - Ratio of correctly predicted positive observations to the total predicted positive observations.
- Recall (Sensitivity) - Ratio of correctly predicted positive observations to the all observations in actual class - yes.
- F1 score - Weighted average of Precision and Recall.
- Accuracy - Ratio of correctly predicted observation to the total observations.

B. K Nearest Neighbour

KNN Algorithm stores all the available date and classifies a new data point based on similar categories.

KNN Distance Measures: Minkowski distance is the genralized distance metric to calculate the distance between two data points.
From the calculated Eucledian distance, the nearest neighbours are known, and are mapped to the similar categories that they belong.

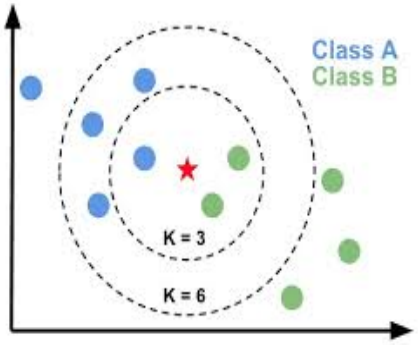

Fig. 9. KNN

C. Naive Bayes

Naive Bayes is a probability-based Machine Learning model which depends on conditional probability.t handles both continuous and discrete data.It is fast and can be used to make real-time predictions. It is not sensitive to irrelevant features. For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter.

Fig. 10. Bayes Theorem

D. Decision Tree

It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.Attribute Selection Measure i.e., information gain has been used which is helping the model to determine which feature gives the maximum information about a class

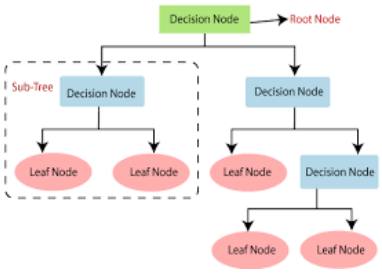

Fig. 11. Decision Tree

In the above figure the decision node is the root node which

reprsents the whole dataset, the leaf node reprents the final node after which the tree cannot be segregated and the sub tree after splitting is done on various conditions.

VII. Results and Conclusions

After the required preprocessing is done, the data is converted to an operational form. 75 percent of the data is used for training the model whereas the remaining 25 percent was used for testing. Four models were applied to the data.

A. Confusion Matrices

1) Logistic Regression: The true positive value from this model is 76.04 percent which shows the 76 times the model has predicted a positive result and false negative value of 4.64 percent shows the correct interpretation of the false data.

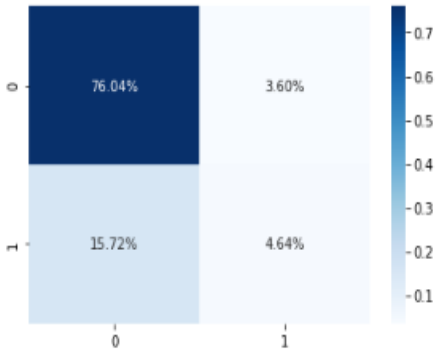

Fig. 12. Logistic Regression- Confusion Matrix

2) KNN: The true positive value from this model is 74.44 percent which shows the 74 times the model has predicted that a customer will churn which is same as the actual output and false negative value of 8.44 percent shows the correct interpretation that the customers are not most likley to leave as predicted by the model.

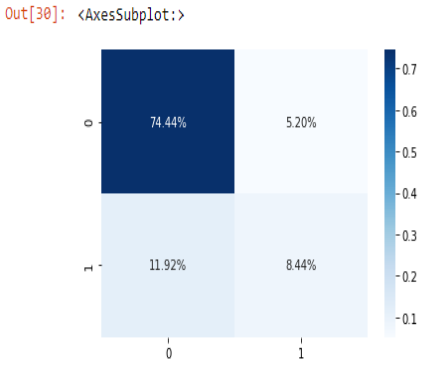

Fig. 13. KNN- Confusion Matrix

3) Naive Bayes: The true positive value from this model is 1861 shows the number of samples with probability close to 1 where the predicted output is closed to the actual output.

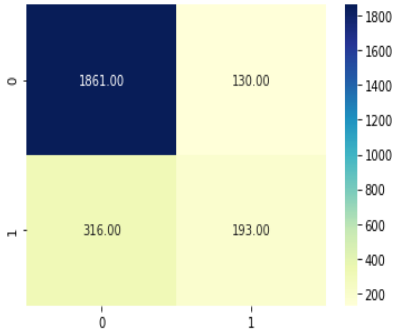

Fig. 14. Naive Bayes- Confusion Matrix

4) Decision Tree: The true positive value from this model is 68.72 percent where the predicted output was 68 perecent accurate as the actual output.

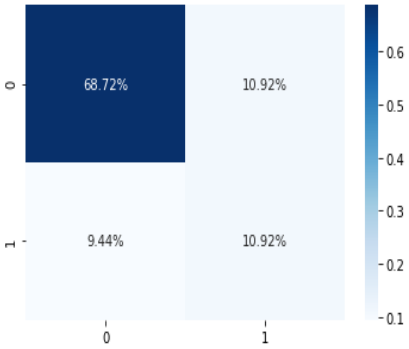

Fig. 15. Decision Tree- Confusion Matrix

B. Accuracy Results

According to the accuracy achieved using the above-mentioned models, the KNN and Naive Bayes models seem to perform the best for predicting bank customer churn.

	Model	accuracy_percentage
0	Logistic Regression	80.68
1	KNN	82.88
2	Decision Tree	79.64
3	Naive Bayes	82.16

Fig. 16. Accuracy

VIII. Conclusion

Customer engagement is one of the primary concerns in the Banking sector. It is very important for banks to forecast the client churning at early stages. Building an early-prediction model to predict the status of a bank's customers may help them to avoid losing their customers as gaining new customers is a costlier affair as compared to retention of present customers. The customer churning rate is the percentage of a bank's customers who try to leave the service. It is imperative to develop a model that can make predictions by using minimal data and provide maximum accuracy. For accurate predictions the selection of independent variables is an important step. This step can be implemented using the feature-selection models. In this study four models were created using only a small amount of data of 10000 samples. Out of the four models two outperformed the others. The KNN model provided an accuracy of 82.88 percent whereas Naïve Bayes model provided an accuracy of 82.16 percent

References

[1] https://ieeexplore.ieee.org/abstract/document/9297529/

[2] https://www.sciencedirect.com/science/article/pii/S0957417408004326

[3] https://ieeexplore.ieee.org/abstract/document/8935884

[4] https://link.springer.com/article/10.1186/s40854-016-0029-6

[5] https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62

[6] $348094541_{M}achine_{L}earning_{B}ased_{C}ustomer_{C}hurn_{P}rediction_{I}n_{B}anking$