A Comparative Study using Feature Selection to Predict the Behaviour of Bank Customers

Sreethi Musunuru[1], *Mahaalakshmi* Mukkamala[1], *Latha* Kunaparaju[2], and *N V Ganapathi* Raju[3]

[1]Dept. of Information Technology, Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, India
[2]Dept. of H and S, Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, India
[3]Dept. of Information Technology, Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, India

Abstract. Though banks hold an abundance of data on their customers in general, it is not unusual for them to track the actions of the creditors regularly to improve the services they offer to them and understand why a lot of them choose to exit and shift to other banks. Analyzing customer behavior can be highly beneficial to the banks as they can reach out to their customers on a personal level and develop a business model that will improve the pricing structure, communication, advertising, and benefits for their customers and themselves. Features like the amount a customer credits every month, his salary per annum, the gender of the customer, etc. are used to classify them using machine learning algorithms like K Neighbors Classifier and Random Forest Classifier. On classifying the customers, banks can get an idea of who will be continuing with them and who will be leaving them in the near future. Our study determines to remove the features that are independent but are not influential to determine the status of the customers in the future without the loss of accuracy and to improve the model to see if this will also increase the accuracy of the results.

1 INTRODUCTION

Machine Learning is highly regarded as a methodology to build business models. Finance and banking require a solid plan to run a business and keep their customers satisfied. A slight overlook could change the dynamic of the business.

The data that is generated on a regular basis from these businesses are in fact very significant for their future growth. In order to get a convincing result, a dataset of bank customers has been collected for model training and testing. Though the dataset collected is close to real world values, it consists of various features that might not actually contribute to the outcome of our study.

The outcome can determine the future relation of the bank customers with the banks; therefore, the methods should produce results with a great accuracy. For this to be possible, we are going to apply a methodology called feature selection that will help us determine the features in the dataset that hold the highest importance in giving the most accurate result. Feature selection is regarded as one of the most important steps in removing unnecessary data and reducing the complexity of the dataset.

This study aims to improve the accuracy of determining which customer will be staying with the bank and who will be leaving using the independent variables provided in the dataset and by removing the independent, non-influential variables that have no significance in affecting the accuracy. By removing these variables, the dataset can be optimised without any loss of accuracy.

2 RELATED WORK

S. Khalid et.al [1]: They explored different dimensionality reduction techniques to understand the strengths and weaknesses of various currently under use procedures. *Chih-Ta* Lin et.al [2]: They used feature selection to convert high-dimensional feature vectors into low-dimensional feature vectors using a real-world dataset. *Lawrence* B. Holder et.al [3]: They explored the improvements in feature extraction methods by working on the current and future trends in feature selection for classification problems. *C. A.* Murthy et.al [4]: He worked on developing a unified framework of feature selection and extraction for both supervised and unsupervised cases. *Zena* M. Hira et.al [5]: They have used feature selection an extraction method to analyse microarrays as they are very large in size. They have also provided a comparison of various selection methods. *Kratarth* Goel et.al [6]: They have used CDFs to improve the accuracy of classification problems. They have provided great results by using the methods on two totally different problem statements. Our study on feature selection is also related to various studies done by other researchers [7...13].

3 METHODOLOGY

1. Data Collection: The dataset for determining customers' behaviour has been collected from Udemy. The dataset has 10,000 rows and 14 columns. The features of this dataset include RowNumber, CustomerId, Surname, CreditScore, Geography, Gender, Age, Tenure, Balance, NumOfProducts, HasCrCard, IsActiveMember, EstimatedSalary and Exited. The study will determine if these features influence the result or if they are just insignificant.

lumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
1	15634602	Hargrave	619	France	Female	42	2	0.00	1	1	1	101348.88	1
2	15647311	Hill	608	Spain	Female	41	1	83807.86	1	0	1	112542.58	0
3	15619304	Onio	502	France	Female	42	8	159660.80	3	1	0	113931.57	1
4	15701354	Boni	699	France	Female	39	1	0.00	2	0	0	93826.63	0
5	15737888	Mitchell	850	Spain	Female	43	2	125510.82	1	1	1	79084.10	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
9996	15606229	Obijiaku	771	France	Male	39	5	0.00	2	1	0	96270.64	0
9997	15569892	Johnstone	516	France	Male	35	10	57369.61	1	1	1	101699.77	0
9998	15584532	Liu	709	France	Female	36	7	0.00	1	0	1	42085.58	1
9999	15682355	Sabbatini	772	Germany	Male	42	3	75075.31	2	1	0	92888.52	1
10000	15628319	Walker	792	France	Female	28	4	130142.79	1	1	0	38190.78	0

Fig. 1. Dataset snapshot

2. Data Cleaning and Pre-processing: The dataset used was very close to the real-world values, therefore, consisted of missing values that could affect the accuracy of the result. The dataset also had to be normalised to convert all the values to be a part of the same range.
For this purpose, KNN had been used to handle these missing values as it can accommodate them near their closest neighbours basing on the other attributes.
For the testing dataset, the value of n was taken as 5 which was later used to calculate the value of k, i.e. k = sqrt(n)/2 which is the Euclidean distance for data plotting whereas the median is considered as the aggregation method.
3. Encoding: As the dataset also contained categorical and text values, encoding had to be done to convert them into numerical inputs. This was achieved using LabelEncoder and OneHotEncoder of the SciKit python library.
4. The methodology of our study next required us to select the best machine learning algorithms that would generate the highest accuracy in predicting which bank customer will be staying and who will be exiting.
Model Selection: Model Selection is quite an important phase in Applying Machine Learning methods to any dataset as the outcome will depend on how effectively the algorithms that have been selected fit the data. Data scientists and analysts use different Machine Learning algorithms on their datasets. We can divide those algorithms into supervised and unsupervised algorithms. Depending on the output label supervised is again classified into classification and regression.

The output label has been identified as classification. Therefore, the following classification algorithms are applied to the dataset:

- K Neighbors classifier: This method had been used as it considers all the data points and classifies them based on the Euclidean distance function.
- Random Forest Classifier: This method had been considered as it uses decision trees for classifying the data. This, in fact, proved to be efficient in working with test errors.
- Extra Trees Classifier: This method was considered as it avoids overfitting the data by randomizing a few decisions and the data in subsets.
- Random forest regressor: A random forest can perform both regression and classification tasks on the data by building multiple decision trees. This has been considered as it offers efficient estimates of test error.

These methods had been used to fit the data points by prediction. A confusion matrix was visualized to understand the errors generated in the process.
The accuracy of the classifiers was tested to determine the best classifier.
5. Feature Selection: It was identified that the accuracy of the prediction can be improved by removing insignificant features as they increased the size of the dataset and increased its complexity. To help the dataset to train faster, feature selection was done.
Extra Trees Classifier was used to visualize the importance of plotting each feature in the dataset.
On visualizing this, the features that are significant were kept and the insignificant features were removed.
Finally, Random Forest classifier was used to plot decision trees to see if there was any loss or improvement in the accuracy after feature selection.

4 RESULTS

The confusion matrix visualized to get the test errors is as follows.

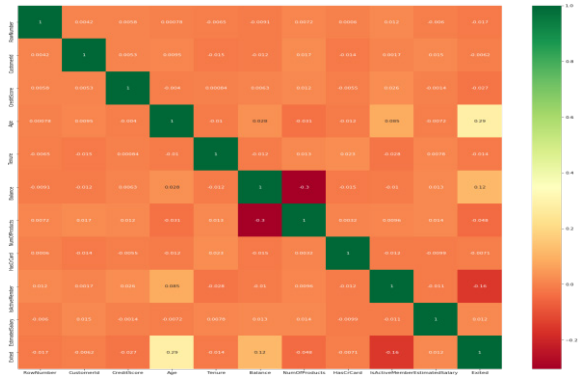

Fig. 2. Confusion matrix

The accuracy score of predicting which customer will be staying and who will be leaving determined using the K Neighbors on the testing dataset was 82.85%.

On plotting a chart using ExtraTrees Classifier, features that were insignificant were removed.

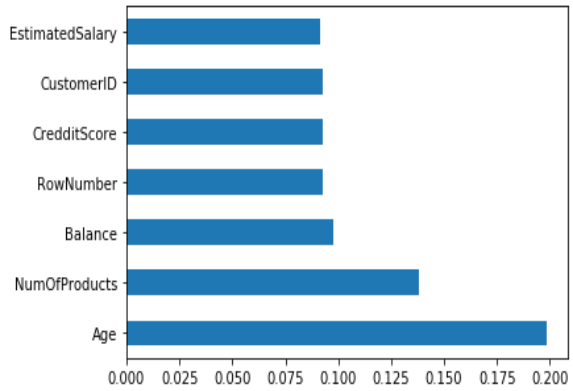

Fig. 3. Bar Chart for features' importance

Significant features determined using this method are EstimatedSalary, CustomerID, CreditScore, RowNumber, Balance, NumOfProducts and Age.

```
In [61]: chi_feature

Out[61]: ['RowNumber',
          'Geography',
          'Gender',
          'Age',
          'Balance',
          'NumOfProducts',
          'IsActiveMember']
```

Fig. 4. Significant features determined using chi-square

The features that have significant importance determined using Chi-square are RowNumber, Geography, Gender, Age, Balance, NumOfProducts, IsActiveMember.

Other features have been removed before plotting and calculating the accuracy score using the training dataset.

On predicting the values using Random Forest Classifier, the accuracy score has improved as follows:

Accuracy of the prediction before feature selection:

Table 2: Accuracy using KNN

CLASSIFIER	ACCURACY
K Neighbors	82.85%

Accuracy of the prediction after feature selection using features determined by ExtraTrees Classifier:

Table 3: Accuracy using Random Forest

CLASSIFIER	ACCURACY
Random Forest	84.55%

Accuracy of the prediction after feature selection using features determined by Chi-square:

Table 4: Accuracy after Feature Selection

CLASSIFIER	ACCURACY
Random Forest	85.15%

To understand the false positive rate and the true positive rate an ROC curve has been visualised below. The ROC curve plots the distributions of the probabilities. The x-axis takes the false positive rate whereas the y-axis takes the true positive rate.

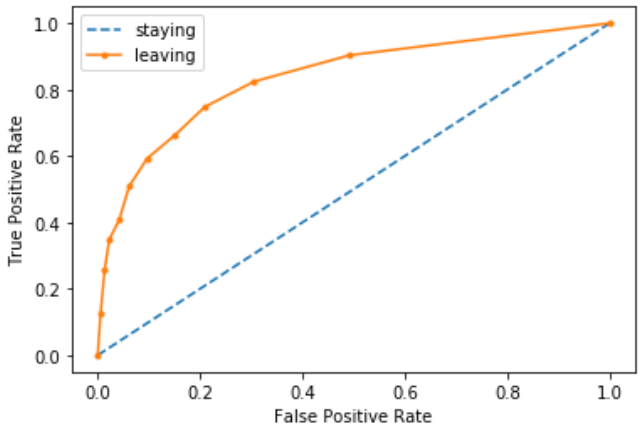

Fig. 5. ROC Curve

5 CONCLUSION

On comparing the accuracy before and after feature selection, it was identified that it increased by 2.3%.
The increase in the accuracy of predicting the 'Exited' feature of the customers' behaviour dataset has helped in improving the business model of the bank.
Bankers can now develop more personalised strategies to get their customers to stay with them and also devise future plans to attract more customers to their banks.
For business analysts or data scientists who will work on such business models, it can be concluded that by feature selection the accuracy cannot just be improved but also decrease the complexity of the dataset and train the dataset faster.
As we have seen through the methods applied above, we can't figure out the best parameters related to the dataset therefore, the future work can involve selecting the number of best parameters to increase the accuracy of the problem statement.

References

1. S. Khalid, T. Khalil, S. Nasreen, *A survey of feature selection and feature extraction techniques in machine learning* (Science and Information Conference, 2014)
2. Chih-Ta Lin, Nai-Jian Wang, Han Xiao, Claudia Eckert, *A survey of feature selection and feature extraction techniques in machine learning*, Journal Of Information Science And Engineering **31**, 965-992 (2015)
3. Lawrence B. Holder, Ingrid Russell, Zdravko Markov, Anthony G. Pipe, B. Carse, *Current And Future Trends In Feature Selection And Extraction For Classification Problems*, International Journal of Pattern Recognition and Artificial Intelligence **19**, No. 02, pp. 133-142 (2005)

4. C. A. Murthy, *Bridging Feature Selection and Extraction: Compound Feature Generation*, IEEE Transactions on Knowledge and Data Engineering Volume: **29** , Issue: 4 , April 1 (2017)
5. Zena M. Hira; Duncan F. Gillies, *A Novel Feature Selection and Extraction Technique for Classification* 2014 14th International Conference on Frontiers in Handwriting Recognition (2015)
6. Kratarth Goel; Raunaq Vohra; Ainesh Bakshi, *A Novel Feature Selection and Extraction Technique for Classification*, 2014 14th International Conference on Frontiers in Handwriting Recognition (ICFHR)
7. N. Chumerin and V. Hulle, M. M, *Comparison of Two Feature Extraction Methods Based on Maximization of Mutual Information* In: Proceedings of the 16th IEEE Signal Processing Society Workshop on Machine Learning for Signal Processing, pp. 343-348 (2006).
8. A. G. K. Janecek and G. F. Gansterer et al, *On the Relationship between Feature Selection and Classification Accuracy*, In: Proceeding of New Challenges for Feature Selection, pp. 40-105 (2008).
9. Yogeswara Reddy B, Srinivas Rao J, Suresh Kumar T, Nagarjuna A, International Journal of Innovative Technology and Exploring Engineering, 8(11), 2019
10. Prasanna Lakshmi, K., Reddy, C.R.K. *A survey on different trends in Data Streams* (2010) ICNIT 2010 - 2010 International Conference on Networking and Information Technology, art. no. 5508473, pp. 451-455. Cited 15 times. 2-s2.0-77955591448 Document Type: Conference Paper Publication Stage: Final Source: Scopus (2010)
11. Padmavathi, K., Sri Ramakrishna, *K. Classification of ECG signal during Atrial Fibrillation using Autoregressive modeling* (2015) Procedia Computer Science, **46**, pp. 53-59. Cited 29 times. 2-s2.0-84931436900 Document Type: Conference Paper Publication Stage: Final Source: Scopus (2015)
12. Dhanalaxmi, B., Apparao Naidu, G., Anuradha, *K. Adaptive PSO based association rule mining technique for software defect classification using ANN* (2015) Procedia Computer Science, **46**, pp. 432-442. Cited 7 times. 2-s2.0-84931310946 Document Type: Conference Paper Publication Stage: Final Source: Scopus (2015)
13. Y.Jeevan Nagendra Kumar, B. Mani Sai, Varagiri Shailaja, Singanamalli Renuka, Bharathi Panduri, *Python NLTK Sentiment Inspection using NaïveBayes Classifier IJRTE* ISSN: 2277-3878, Volume-**8**, Issue-2S11, September (2019)