"Bank Customer Churn Prediction: A Comparative Analysis Using Machine Learning Classification Models"

Manjil Nepal · Sunit Soni ·Jayash Shrestha

Department of Computer Science and Engineering, SRM University-AP Andhra Pradesh

Abstract

Customer churn, a prevalent concern for banks worldwide, involves the departure of customers from their banking services. Predicting churn necessitates meticulous data pre-processing, feature analysis, and model training, incorporating techniques like logistic regression, support vector machines, and ensemble methods. By evaluating model performance using metrics such as confusion matrices and AUC curves, banks can pre-emptively address customer attrition and implement tailored retention strategies to bolster customer loyalty and satisfaction.

Keywords: Customer Churn Prediction · Machine Learning · Support Vector Machine · Nearest Neighbors · Decision Tree · Random Forest

1. Introduction

Customer churn, a prevalent concern for banks worldwide, involves the departure of customers from their banking services. Predicting churn necessitates meticulous data pre-processing, feature analysis, and model training, incorporating techniques like logistic regression, support vector machines, and ensemble methods. By evaluating model performance using metrics such as confusion matrices and AUC curves, banks can pre-emptively address customer attrition and implement tailored retention strategies to bolster customer loyalty and satisfaction.

1.1 Problem Description

The problem at hand involves predicting customer churn for a U.S. bank using machine learning algorithms. The objective is to develop predictive models using Support Vector Machines (SVM), k-Nearest Neighbors (KNN), and Decision Trees to classify customers as churners or non-churners based on various features such as banking activities, transaction history, etc. By comparing the performance of these models using evaluation metrics like accuracy, precision, recall, and F1-score, the aim is to identify the most effective approach for accurately predicting customer churn. This predictive capability will enable the bank to proactively implement targeted retention strategies, thereby minimizing customer attrition and improving overall business performance.

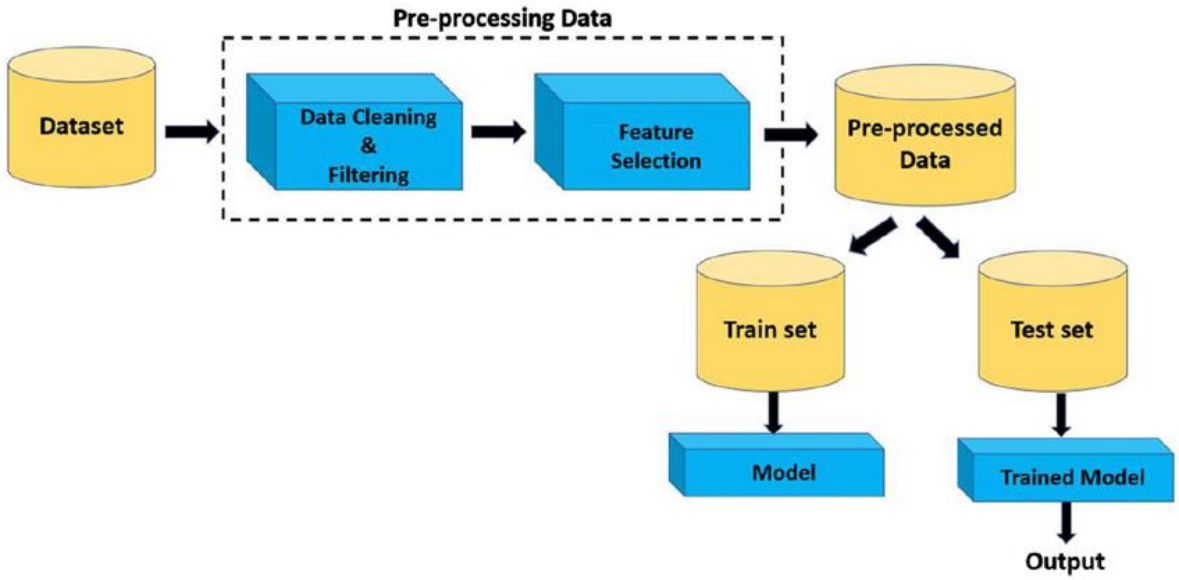
1.2 Author's Contribution

Summary of our contribution is as follows:

- We have applied gravitational search algorithm to perform feature selection and
  to reduce the dimensions of the dataset.
- After, pre-processing of data, we have applied some of the famous machine learning techniques which are used for predictions like logistic regression, SVM, etc and k-fold cross validation has been performed to prevent overfitting.
- Then we have used the power of ensemble learning to optimize algorithms and achieve better results.
- Then we have evaluated the algorithms on test set using confusion matrix and AUC curve, which have been mentioned in form of graphs and tables to compare which algorithm performs best for this dataset.

2. Dataset

2.1 Pre-processing



2.2 Initial Dataset

RowNo	CustomerId	Surname	CreditScore	Country	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

rows × 14 columns

2.3 Feature vector and their types

Attribute	Value
Age	Numeric
Tenure	Numeric
Balance	Numeric
NumOfProducts	Numeric
HasCrCard	Binary
IsActiveMember	Binary

2.4 Feature Scaled Dataset

	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember
0	0.293517	-1.041760	-1.225848	-0.911583	0.646092	0.970243
1	0.198164	-1.387538	0.117350	-0.911583	-1.547768	0.970243
2	0.293517	1.032908	1.333053	2.527057	0.646092	-1.030670
3	0.007457	-1.387538	-1.225848	0.807737	-1.547768	-1.030670
4	0.388871	-1.041760	0.785728	-0.911583	0.646092	0.970243

3. Machine Learning Models Used

3.1 Support Vector Machine

Support Vector Machine (SVM) is a powerful supervised learning algorithm used for classification and regression tasks. Its main objective is to find the hyperplane that best separates classes in the feature space. SVM aims to maximize the margin between the classes, making it robust to outliers. It can handle both linear and non-linear data through the use of different kernel functions such as linear, polynomial, and radial basis function (RBF). SVM is effective in high-dimensional spaces and is widely used in various fields such as image classification, text categorization, and bioinformatics.

3.2 K-Nearest Neighbors

K-Nearest Neighbors (KNN) is a simple yet effective algorithm used for classification and regression tasks. In KNN, the class of a data point is determined by the majority class among its K nearest neighbors in the feature space. It is a non-parametric method, meaning it does not make any assumptions about the underlying data distribution. KNN is easy to understand and implement, but it can be computationally expensive, especially with large datasets. It is sensitive to the choice of distance metric and the value of K.

3.3 Decision Tree

Decision Tree is a versatile supervised learning algorithm used for classification and regression tasks. It works by recursively partitioning the feature space into smaller subsets based on the values of features, with the goal of maximizing the homogeneity of the target variable within each subset. Decision trees are interpretable and can handle both numerical and categorical data. However, they are prone to overfitting, especially when the tree depth is not properly controlled. Techniques like pruning and ensemble methods such as Random Forest and Gradient Boosting are often used to mitigate overfitting and improve performance.

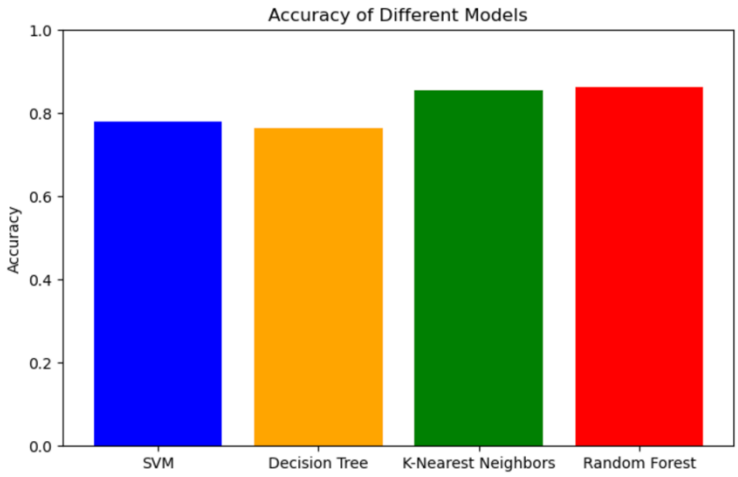
3.4 Random Forest

Random Forest is an ensemble learning method that operates by constructing multiple decision trees during training and outputting the class that is the mode of the classes of the individual trees. It builds each tree in the ensemble using a random subset of the features and a random subset of the training data. This randomness helps to reduce overfitting and increase the robustness of the model. Random Forest is widely used for classification and regression tasks due to its simplicity, scalability, and ability to handle high-dimensional data. It is less prone

to overfitting compared to individual decision trees and often yields higher accuracy.

4. Performance analysis

```
--------------------------------------------------------------------------------

Support Vector Machine:
Best parameters: {'C': 1, 'gamma': 'auto', 'kernel': 'rbf'}
Accuracy: 78.10%
Classification Report:
              precision    recall  f1-score   support

           0       0.92      0.80      0.85      2419
           1       0.46      0.72      0.56       581

    accuracy                           0.78      3000
   macro avg       0.69      0.76      0.71      3000
weighted avg       0.83      0.78      0.80      3000

--------------------------------------------------------------------------------


Decision Tree:
Best parameters: {'max_depth': 5}
Accuracy: 76.43%
Classification Report:
              precision    recall  f1-score   support

           0       0.92      0.78      0.84      2419
           1       0.43      0.70      0.54       581

    accuracy                           0.76      3000
   macro avg       0.67      0.74      0.69      3000
weighted avg       0.82      0.76      0.78      3000

--------------------------------------------------------------------------------


K-Nearest Neighbors:
Best parameters: {'n_neighbors': 9, 'weights': 'uniform'}
Accuracy: 85.57%
Classification Report:
              precision    recall  f1-score   support

           0       0.88      0.95      0.91      2419
           1       0.70      0.45      0.55       581

    accuracy                           0.86      3000
   macro avg       0.79      0.70      0.73      3000
weighted avg       0.84      0.86      0.84      3000

--------------------------------------------------------------------------------


Random Forest:
Best parameters: {'max_depth': 10, 'min_samples_split': 5, 'n_estimators': 30}
Accuracy: 86.27%
Classification Report:
              precision    recall  f1-score   support

           0       0.88      0.97      0.92      2419
           1       0.75      0.44      0.55       581

    accuracy                           0.86      3000
   macro avg       0.81      0.70      0.74      3000
weighted avg       0.85      0.86      0.85      3000

--------------------------------------------------------------------------------
```



5. Conclusion and future findings

The study highlights the importance of machine learning in predicting bank customer churn. Through comparative analysis, Random Forest emerges as the most accurate model, achieving an 86.77% accuracy rate. These results provide valuable insights for banks to implement proactive retention strategies, thereby fostering stronger customer relationships and enhancing business performance. Moving forward, further research could delve into advanced techniques to refine churn prediction accuracy and monitor the effectiveness of implemented strategies for continuous improvement in customer relationship management practices.

6. References

1. Vafeiadis, T., Diamantaras, K. I., Sarigiannidis, G., & Chatzisavvas, K. C. (2015). A comparison of machine learning techniques for customer churn prediction. Simulation Modelling Practice and Theory, 55, 1-9.
2. Lalwani, P., Mishra, M. K., Chadha, J. S., & Sethi, P. (2022). Customer churn prediction system: a machine learning approach. Computing, 104(2), 271-294.
3. Qureshi, S. A., Rehman, A. S., Qamar, A. M., Kamal, A., & Rehman, A. (2013, September). Telecommunication subscribers' churn prediction model using machine learning. In Eighth international conference on digital information management (ICDIM 2013) (pp. 131-136). IEEE.
4. Geiler, L., Affeldt, S., & Nadif, M. (2022). A survey on machine learning methods for churn prediction. International Journal of Data Science and Analytics, 14(3), 217-242.
5. Rahman, M., & Kumar, V. (2020, November). Machine learning based customer churn prediction in banking. In 2020 4th international conference on electronics, communication and aerospace technology (ICECA) (pp. 1196-1201). IEEE.
6. Khodabandehlou, S., & Zivari Rahman, M. (2017). Comparison of supervised machine learning techniques for customer churn prediction based on analysis of customer behavior. Journal of Systems and Information Technology, 19(1/2), 65-93.

7. Sisodia, D. S., Vishwakarma, S., & Pujahari, A. (2017, November). Evaluation of machine learning models for employee churn prediction. In 2017 international conference on inventive computing and informatics (icici) (pp. 1016-1020). IEEE.
8. Gaur, A., & Dubey, R. (2018, December). Predicting customer churn prediction in telecom sector using various machine learning techniques. In 2018 International Conference on Advanced Computation and Telecommunication (ICACAT) (pp. 1-5). Ieee.
9. Erdem, Z. U., Çalış, B., & Fırat, S. Ü. (2021). Customer Churn prediction analysis in a telecommunication company with machine learning algorithms. Endüstri Mühendisliği, 32(3), 496-512.
10. Rahman, M., & Kumar, V. (2020, November). Machine learning based customer churn prediction in banking. In 2020 4th international conference on electronics, communication and aerospace technology (ICECA) (pp. 1196-1201). IEEE.