**Bank Churn Prediction**

**Abstract**

This study aims to make a predictive model through which we are able to find out the customer churn rate who are going to discontinue the bank services. For maximizing the revenue of any organization customer churn is one of the most important factors. These days banks in the financial industry facing huge competition and if they are not able to keep the existing customer then there is a negative impact on their revenue as well as they also lose their reputation in the market. So, they need to focus on reducing loyal customer churn rates since the probability of selling any service to an existing or loyal customer is a lot higher than to a new prospective customer. Now in this research, we build the predictive model with the help of the Machine Learning Algorithm and since this is the binary classification task as we need to make prediction between Exited (1) and Retained (0) then we are going to advance the classification algorithms like Random Forest, Decision Tree, and XGB classifier. The dataset is taken from kaggle and the dataset contains 10,000 rows (samples) and 14 columns (features).

**Keywords**

Classification, Exited, Retention, Banking, Financial Industry, Decision Tree, Random Forest, and Extra Gradient Boosting.

**Introduction**

In this new era of technology where everything is online and because of the digital growth of the society, maintaining and managing the relationship with the customer is very essential to generate the positive revenue, in terms of business this concept is called “Customer Relationship Management” (CRM). In this thesis, we are not only trying to make predictive model but along with this we try to get different insights from dataset using Exploratory Data Analysis (EDA). The dataset is taken from kaggle which is related to 3 geographical location France, Germany, and Spain. It has 10,000 rows with 14 columns such as age, gender, balance, tenure and exited column to indicate whether the customer has left the bank or not which has been as a target variable. Dataset indicates that 20% of the customers who are loyal to the bank for more than 3 years left the bank and there is no missing data found during data preparation. And at last in this thesis we build the classification model using Machine Learning Algorithm such as decision tree, extra gradient boosting and random forest to predict which customer is likely to churn in future.

**Literature Review**

**Himani Jain, Garima Yadav, R. Manoov** proposed the idea of Machine Learning Technique for churn prediction and retention as it is the biggest deciding factor which directly put high impact on revenue of any organization. In this paper, they build the predictive model for customer churn so that the proper strategy can be built by the organization through Exploratory Data Analysis (EDA) to retain their loyal customer. In this paper they used 4 machine learning algorithms logistic regression, support vector machine (SVM), random forest and XGBoost for modelling.

**Isabelle Tandan & Erika Goteman** discuss about different classification algorithm for making Bank Churn Prediction model, by using random forest, logistic regression and k-nearest neighbor (KNN) and then they validate the performance of model using k-fold cross validation as it is having a computational advantage.

[**Abbas Keramati**](https://jfin-swufe.springeropen.com/articles/10.1186/s40854-016-0029-6#auth-Abbas-Keramati)**,**[**Hajar Ghaneei**](https://jfin-swufe.springeropen.com/articles/10.1186/s40854-016-0029-6#auth-Hajar-Ghaneei)**& [Seyed Mohammad Mirmohammadi](https://jfin-swufe.springeropen.com/articles/10.1186/s40854-016-0029-6" \l "auth-Seyed_Mohammad-Mirmohammadi)**presented the idea of predictive model for customer churn as in this competitive atmosphere it’s become very difficult for the organization to retain their loyal customers. Based on the existing information they collect the data from organization, apply Exploratory Data Analysis (EDA) for analysis and then apply decision tree algorithm to build the predictive model.

**Benlan Hea,b, Yong Shic, Qian Wand, Xi Zhaoc** propose the SVM model for the prediction of customer attrition in the commercial banks. In their research they took an example Chinese commercial banks where they need to avoid the loss of customers in order to improve their profits and capital. They build a predictive model by using the support vector machine (SVM) and since the dataset is quite imbalanced so they do random sampling in order to enhance the performance of model.

**Methodology**

The methodology involved the use of Python environment for programming, followed by Exploratory Data Analysis (EDA) to obtain valuable insights from the data for the organization's retention strategy. Feature Engineering was then applied to enhance the prediction model's performance. The data was split into training and testing sets, and a supervised machine learning algorithm was utilized to develop the prediction model. Finally, the accuracy score and confusion matrix were utilized to assess the prediction model's performance. In summary, this is the methodology employed in the thesis, and each step will be discussed briefly..

**Exploratory Data Analysis (EDA)**

In order to create a retention strategy, various analyses were conducted based on the available data, which are outlined below:

- The dataset is slightly imbalanced, with only 20% of customers having exited and 80% of customers retained.

- Germany had the highest number of churned customers, while France had a higher retention rate.

- Male customers were found to have a higher retention rate compared to female customers.

- Customers who had a credit card were more likely to churn.

- Inactive members were found to have a higher churn rate.

- There was no difference in credit score distribution between the customers who churned and those who were retained.

- Surprisingly, older customers had a higher churn rate compared to younger ones, so the bank needs to develop a strategy to retain their older customers.

- Extremely low or high tenure periods were found to be associated with a higher churn rate.

- A concerning observation was that customers with higher account balances were more likely to churn, which could lead to a capital deficiency in the bank.

- Estimated salary and number of products had no impact on customer churn.

- There was no significant intercorrelation between the features, so multicollinearity was not a concern.

**Feature Engineering**

In the process of feature engineering, two types of features, namely numerical and categorical, are generated. Subsequently, irrelevant columns such as Row Number, Customer ID, and Surname are eliminated. Categorical data is then converted into numerical data through the application of one-hot encoding. Next, the min-max scalar technique is implemented to scale down all feature values between 0 and 1. The prepared dataset is then segregated into two variables, one containing all the features, and the other containing the target column. Finally, the dataset is split into two sets, with 80% of the data being used for training and the remaining 20% for testing. Two baseline models are subsequently developed based on this process.

**Algorithms**

**Logistic Regression** is a classification algorithm that estimates the probability of an instance belonging to a particular class. The sigmoid function is used to produce a value between 0 and 1 for this estimation. To predict the class of an instance, a threshold value is set based on the estimated probability.

**Random Forest** is an ensemble technique that uses bagging. It builds multiple decision trees on random samples, hence the name "Random Forest". The technique combines decision trees with row and column sampling to reduce variance, using deep decision trees as the base models. By reducing variance during aggregation, the best model is obtained.

**XGBoost** (Extreme Gradient Boosting) is a boosting ensemble technique that sequentially trains a group of weak learners. After each training round, the error is computed and parameters such as weights and biases are tuned to reduce it. The updated parameters are then used to train the next model until the best model with low bias and low variance is obtained. XGBoost is an improvement over gradient boosting and significantly boosts model performance while being computationally efficient.

In cases where the dataset is imbalanced, **SMOTE** (Synthetic Minority Oversampling Technique) can be used to balance it. SMOTE is an oversampling technique that generates synthetic samples from the minority class and reduces the risk of overfitting. Undersampling can lead to loss of information, making SMOTE a more effective technique.

**Results Analysis**

- Logistic Regression accuracy: 84.7%; F1-score: 0.56

- Random Forest Classifier accuracy: 85.5%; F1-score: 0.55%

- XGBoost classifier accuracy: 85%; F1-score: 0.57%

- Accuracy of the models is good, but F1-score is poor

- Dataset is balanced using SMOTE oversampling technique

- F1-score of Logistic Regression model after applying SMOTE: 0.96%

- F1-score of Random Forest classifier after applying SMOTE: 1

- F1-score of XGBoost classifier after applying SMOTE: 0.96

- The classification model helps identify customers likely to churn in the future

- Exploratory Data Analysis (EDA) determines attribute contribution to the model

- Random Forest performs the best, with maximum F1-score.

The experiments that we conducted using machine learning algorithms and a classification model produced impressive and convincing results in predicting which customers are likely to churn in the future. Through Exploratory Data Analysis (EDA), we identified which attributes have the most impact on the model and which have less. Although all models performed well, Random Forest stood out with the highest F1-score, indicating its superior performance.

**Conclusion**

The problem at hand cannot be solved solely through statistical modeling, so we have opted for a machine learning algorithm instead. The prediction of customer churn is a complex task due to its temporal nature, which adds to the overall complexity of data analysis. Machine learning algorithms help us understand unknown features of churn prediction and identify which features are more important in making decisions. This information can then be used to provide valuable insights to organizations, enabling them to devise strategies to retain potential customers. Based on our study, we have concluded that the Extreme Gradient Boosting Algorithm (XGB) outperforms all other algorithms in terms of the data. Thus, we can infer that employees have a higher probability of staying with the company and are less likely to leave.

**References**

1. Machine learning for customer churn prediction in telecom industry. (2020). Machine Learning for Customer Churn Prediction in Telecom Industry, 1–10. https://www.sciencedirect.com/science/article/pii/S1877050920301962

2. Predicting customer churn in the banking sector using machine learning techniques. (2020). Predicting Customer Churn in the Banking Sector using Machine Learning Techniques, 1–10. https://www.sciencedirect.com/science/article/pii/S2666827020300311

3. Customer churn prediction for mobile telecommunication service providers using machine learning algorithms. (2020). Customer Churn Prediction for Mobile Telecommunication Service Providers using Machine Learning Algorithms, 1–16. https://link.springer.com/article/10.1007/s41019-020-00127-2

4. A comparative study of machine learning algorithms for predicting customer churn in banking sector. (2019). A Comparative Study of Machine Learning Algorithms for Predicting Customer Churn in Banking Sector, 1–10. https://www.sciencedirect.com/science/article/pii/S1877050919320326

5. A predictive modeling approach to customer churn analysis in the telecommunication industry. (2019). A Predictive Modeling Approach to Customer Churn Analysis in the Telecommunication Industry, 1–10. https://www.sciencedirect.com/science/article/pii/S2405452618310285

6. A hybrid method for customer churn prediction in the banking sector using feature selection and machine learning algorithms. (2021). A Hybrid Method for Customer Churn Prediction in the Banking Sector using Feature Selection and Machine Learning Algorithms, 1–12. https://www.sciencedirect.com/science/article/pii/S2405452621000516

7. Customer churn prediction in the telecom industry using machine learning techniques. (2018). Customer Churn Prediction in the Telecom Industry using Machine Learning Techniques, 1–10. https://www.sciencedirect.com/science/article/pii/S1877050918310963