**Market Target Prediction**

**(BANKING INDUSTRY)**

Ajay Akshat Wadhwa (502104008)

MBA - BABD

LM Thapar School of Management

Priyanshi Sharma (502104175)

MBA - BABD

LM Thapar School of Management

***Abstract*—** ***The challenging problem is the bank's direct marketing effort for supplying items that fulfill consumers' wants. The bank's direct marketing data analysis is critical to its job in predicting whether clients would sign a long-term deposit with it. Banks may benefit from a system that can foresee such clients' demands to improve their marketing campaign methods. Unfortunately, because the available knowledge is asymmetric, it is not easy to anticipate customers' wants. This paper presents an algorithm for analyzing asymmetric information. To increase the accuracy of the prediction, the SMOTE approach is utilized to change the data. The suggested method's performance is assessed and compared to the Decision Tree. All predictors' prediction accuracies are shown in the experimental findings.***

***Keyword - One of the most widely used approaches to overcome the imbalance problem is the synthetic minority oversampling technique (SMOTE). Its goal is to achieve a more balanced distribution of classes by recreating minority class examples at random. SMOTE creates new minority instances by combining existing minorities.***

# INTRODUCTION

Artificial intelligence has advanced in recent years, allowing humans to do previously impossible tasks.

The bank's direct marketing activities are pretty practical for introducing new goods to clients. Data analysis may determine which form of direct bank marketing to use. Viral marketing is a type of marketing that uses the internet and communication technology to reach out to customers directly. Customers can get them by email, regular mail, or phone. Customers may have direct access to the goods, allowing them to choose what they want. On the other hand, if the market fails to match the wants of its clients, the items are likely to be rejected.

During the global financial crisis, banks prioritized risk management, and there has been a continued focus on how risks are detected, measured, reported, and managed.

By finding complicated, nonlinear patterns within large datasets, machine learning, which has been recognized as a technology with substantial implications for risk management, can aid in the building of more accurate risk models. With each new piece of information supplied, these models' predictive capacity can improve, resulting in increased predictive power over time. Machine learning is projected to be used in numerous areas of a bank's risk management department. Machine learning has also been suggested as a project that might aid in the modernization of risk management processes in banks.

Customers are the most valuable assets in any business since they are the primary source of revenue. Companies are becoming increasingly conscious that it takes significant effort to gain new clients and keep existing ones. Churners are people who switch careers for a variety of reasons. To decrease customer turnover, the company must correctly predict customer behavior and identify links between customer attrition and issues under its control.

## *About Problem*

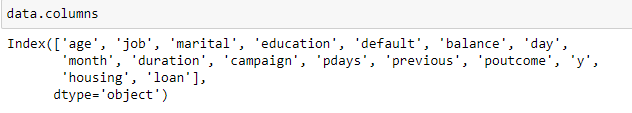
Predicting when a client is likely to churn is a substantial additional potential revenue stream for any company. Aside from the immediate income loss that a customer's departure causes, the costs of acquiring that client may not have been covered by the customer's spending yet. Customer churn prediction is to identify customers who are most likely to leave a company. To keep present customers, the banking industry has to understand the reasons for churn, which may be determined through information collected from data.

## *Motivation*

Boosting algorithms are used to improve a customer churn prediction model by classifying customers into two categories based on how much weight the boosting algorithm provides. A high-risk consumer cluster was detected as a result. Logical regression is used as a primary learner, and a churn prediction model is developed for each set independently. Compared to a single logistic regression model, the experimental results showed that the boosting strategy separates churn data.

# DATASET

The data set used to solve this problem has he information that pertains to a Portuguese financial institution's direct marketing initiatives. Phone calls were used in the marketing activities. Frequently, many contacts with the same client were necessary to determine if the product (bank term deposit) would be subscribed to ('yes') or not ('no') by the consumer.



Output variable (desired target)

*y - has the client subscribed at the end of the campaign? (binary: "yes","no")*

# PROCEDURES AND METHODOLOGIES

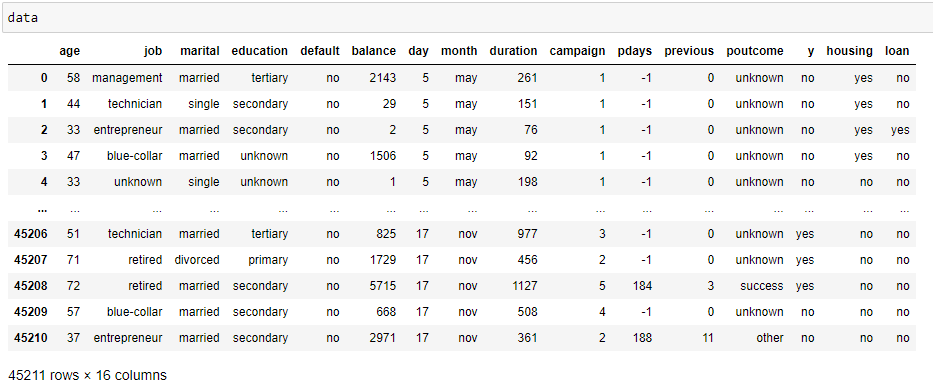
To learn about the conversion of the clients and prediction to enrol with the bank we have used the above mentioned data set. We then applied two Models to depict how accurate the algorithm is.

We used Jupyter Notebook in Anaconda for the programming.

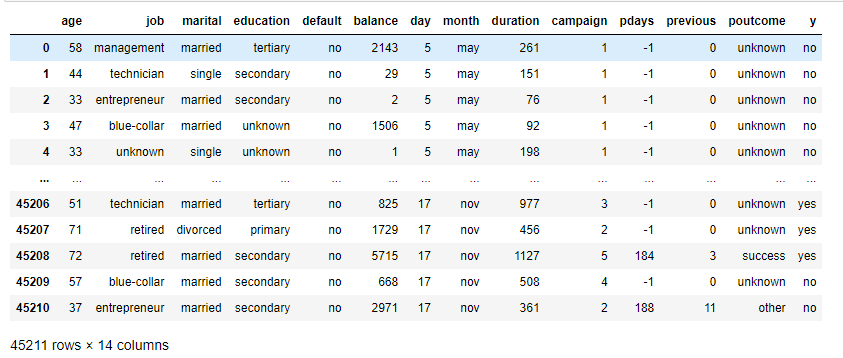
Data Mining and Pre-processing: We first read the data and extracted the columns that are required by us to predict the market conversion. Next we looked for null values or illegal values in the complete data followed by assigning the target variable.

Then we convert categorical data into dummy or indicator variables.

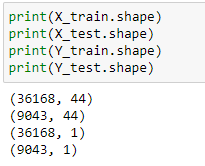
Raw Data Before pre-processing:



Data after preprocessing:



By using model selection we hence split the data for Training and testing followed by Standardization.



After that, we carry out standardisation. Standardization is applied to regularly distributed data values. Furthermore, standardisation tends to make the dataset's mean equal to 0 and the standard deviation equal to 1.

MODELS

1. Naive-Bayes Model

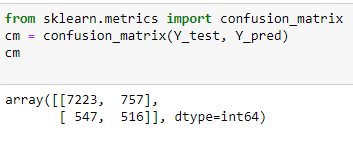
The Nave Bayes method is a supervised learning strategy based on the Bayes theorem for coping with classification tasks.

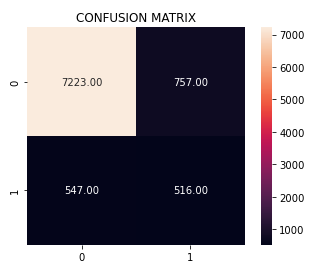
It's mostly used for text classification jobs that require a big training dataset.

The Nave Bayes Classifier is a simple and effective classification approach that assists in the building of fast machine learning techniques that can make precise forecasts.

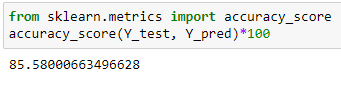
The Bayes hypothesis, often known as Bayes' rule or Bayes' law, is a mathematical formula for calculating the probability of a hypothesis based on prior data. This is governed by conditional probability.

Now Naive-Bayes Model is applied to the training data set and Y-variable is predicted in terms of X test data set. A confusion matrix is then created to check the useful ness of the model.





We now check the accuracy of the model.



The accuracy of the Naive-Bayes Model in case of or dataset for prediction of Market Target is **85.58%.**

1. KNN Model

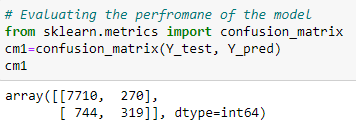
The K-Nearest Neighbour method is based on the Supervised Learning methods and is among the most basic Machine Learning algorithms.

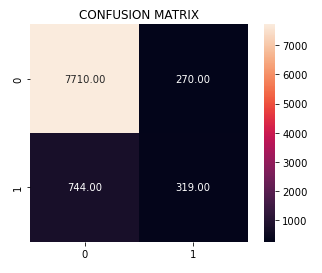
The K-NN algorithm assumes that the new case/data and older cases are comparable and places the new case in the category that is most similar to the previous categories.

The K-NN method saves all data available and classifies a new data point based on its similarity to the current data. This implies that fresh data may be quickly sorted into a well-defined category using the K-NN method.

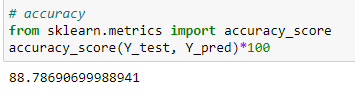
The K-NN algorithm could be used for both regression and classification, however it is more commonly used for classification tasks.

K-NN Model is hence applied to the data set and confusion matrix is created.





Then the accuracy of the model is tested.

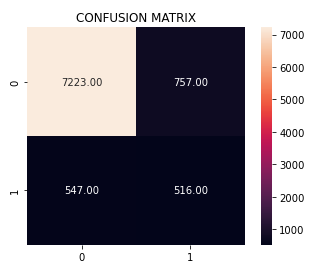
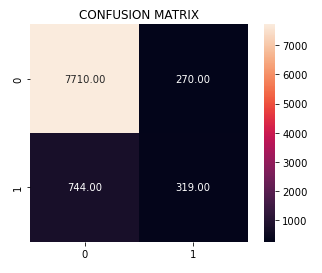


The accuracy of the K-NN Model is 88.78% which is much more than the Naive-Bayes Algorithm.

# RESULTS

|  |  |  |
| --- | --- | --- |
|  | Naive-Bayes | K-NN |
| Accuracy | 85.58% | 88.78% |

# COMPARISION OF CONFUSION MATRICS

Hence we now know that using K-NN Model is more effective when it comes to Bank Marketing Target Prediction.

# RELATED WORK

Ning et al. conducted an experimental study of customer attrition prediction in the telecom industry, arguing that boosting should be used to improve the model. Unlike previous promoting methods that increase a basis learner's accuracy, the author proposed classifying clients into two groups depending on the weight provided by the boosting technique. Consumers with the highest churn proclivity can be engaged in retention efforts in the suggested approach's "Implementation Zone."

To handle the data distribution problem, Y.Xie et al. employed an improved balance random forest (IBFR) model, which is a combination of balanced and weighted random forests. The most remarkable qualities of IBRF are continually taught by altering the class distribution and applying more substantial penalties for misclassification of the minority class. The trials employed a Chinese bank dataset, and the results demonstrated that IBRF outperforms artificial neural networks in terms of accuracy. Artificial neural networks include artificial neural networks, decision trees, and support vector machines.

To predict customer attrition, P.C.Pendharkar suggested two neural networks (NN) models based on Genetic Algorithms (GA). The first GA-based NN model used entropy-based criteria to predict customer turnover, whereas the second GA-based NN model attempted to directly improve customer churn forecast accuracy. They used a real-world customer dataset and three different NNs to evaluate the two GA-based NN models with a statistical Z-score model, leveraging model assessment measures such as prediction accuracy, top 10% docile lift, and area under the Receiver Operating Characteristics (ROC) curve. Both GA-based NN models beat the statistical z-score model on all performance criteria, according to the results of the experiments.

Ver-braken et al. are among those who have contributed to this project. Proposed a new performance indicator called the expected maximum profit criteria, which is tied to the critical goals of end-users. The recommended approach assists enterprises in selecting the classifier that maximizes profit and provides data on the fraction of the customer base that should be targeted for retention.

Ning Lu proposed employing boosting algorithms to enhance a customer churn prediction model by classifying customers into two categories based on the boosting algorithm's weight. A high-risk consumer cluster was detected as a result. Logical regression is used as a primary learner, and a churn prediction model is developed for each set independently. Compared to a single logistic regression model, the experimental results showed that the boosting strategy separates churn data.

## REFERENCES

1. L. Ning, L. Hua, L. Jie, Z. Guangquan, “A customer churn prediction model in telecom industry using boosting”, IEEE Trans. Ind. Inform. 10 (2014) 1659– 1665.
2. Yaya Xie, Xiu Li, E.W.T. Ngai, Weiyun Ying, “Customer churn prediction using improved balanced random forests”, Expert Systems with Applications 36 (2009) 5445–5449.
3. T. Verbraken, W. Verbeke, B. Baesens, “A novel profit-maximizing metric for measuring classification performance of customer churn prediction models”, IEEE Transaction on Knowledge and Data Engineering 25 (2013) 961–973.
4. P.C. Pendharkar, “Genetic algorithm-based neural network approaches for predicting churn in cellular wireless network services,” Expert System Application 36 (2009) 6714–6720.
5. Ning Lu, Hua Lin, Jie Lu, Guangquan Zhang “A Customer Churn Prediction Model in Telecom Industry Using Boosting,” IEEE Transactions on Industrial Informatics, vol. 10, no. 2, May 2014.