CHURN MODELLING

Arun Chilukuri and Sachin Garg*
Data Science
CUNY Graduate Center
Sgarg1@gradcenter.cuny.edu
Achilukuri@gradcenter.cuny.edu

Abstract

In the modern world, with banking and internet becoming so accessible, a lot of people are starting to misuse these banking services just for temporary promotional benefits. Churn prediction is the process of identifying which consumers are most likely to stop using a service or to cancel their subscription. Some practical examples are, Amazon and Linkedin provide a one-month free subscription to their customers in the hope of having continuous relationship with customers, but several customers just cancel subscription once a free period ends. Our motivation in this project to implement all the machine learning models (from linear to tree models to neural networks) that we have learned during the semester.

1 Introduction

Banks want to keep as many current customers as they can. Naturally, they are interested in learning whether the demands of their customers are being satisfied or if any of them have exit plans from the business. The bank can take steps to persuade the customer to stay if it has suspicions that they would favor another business. Our aim in this project to build several machine learning models and deep learning neural networks to predict which customers decide to stay or to leave the bank and which model provides the best accuracy. The inputs are the characteristics of a customer like Age, Gender, Tenure, Balance and other parameters. We then use a logistic regression, decision trees, bagging, boosting, random forest, ANN etc. to predict the output 1, 0, where 1 represents customer leaving the bank and 0 represents customer staying with the bank.

2 Dataset and Features

There are a total of 10,000 observations with 14 columns of bank customer data. We then split the dataset into 80:20 training:testing using train test split method. There are no missing values in the data, but there are few outliers present in the Age variable only. It maybe because in some cases when customer's age is high there is a high chance of account closure due to the customer death. But, we also did Standard Scaler method to encounter the outliers. Then, we drew a correlation matrix heatmap to see what factors are most affecting the output. We build some linear machine learning models like Logistic Regression, Regularization using (Lasso, Ridge and ElasticNet). We build some tree-based models like Decision Trees, Random Forest etc. We also build some ensemble methods like Bagging and Boosting for better model performance. Finally, we build a deep learning Artificial Neural Network (ANN) and compare the accuracy of each model.

*Use footnote for providing further information about author (webpage, alternative address)—*not* for acknowledging funding agencies.

CSCI 74020 - Machine Learning 2022, CUNY, NY. (LateX template borrowed from NIPS 2017.)

3 Methods

a) Logistic Regression: It is a model which predicts the probability of an event taking place by having log-odds for event to be a linear combination of one or more independent variables. The advantages of using logistic regression is that it is easy implement and interpret. It performs good on a simple data. But, disadvantage is that it assumes a linear relationship between dependent and independent variables.

b) Regularization: Regualarization is a technique to reduce model complexity and prevent overfitting. It can be done by adding a penalty term to the OLS formula. There are three techniques for regularization. Ridge, Lasso and ElasticNet.

i) *Ridge (L2 Norm)*: It adds a penalty term lambda to the biased-term square. Large lambda adds too much shrinkage and leads to underfitting.

ii) *Lasso (L1 Norm)*: Similar to Ridge, it also adds a penalty term lambda, but takes absolute value. Large lambda will make more coefficients than should be 0 and cause model underfitting.

iii) *ElasticNet(L1 Ratio)*: It is more flexible regularization between Ridge Lasso. lambda is shared penalization parameter while alpha sets the ratio between L1 and L2 regularization.

c) Decision Trees: It is supervised learning technique used to build predictive machine learning models for both categorical or continuous target variables. The decision trees are easy to interpret and easy to visualize, but the main problem is that they easily overfit because they are biased in cases of class imbalance.

d) Random Forest: It is simply a bootstrapped version of comprised thousands of decision trees. More trees we have, more robust the algorithm will be. It can handle large datasets very efficiently and provides a great accuracy.

e) Bootstrap Aggregation (Bagging): It is one of ensemble method where a subset of data is selected with replacement. It uses the output average prediction which reduces variance and produces more accurate models.

f) Boosting: It is also a ensemble technique to build multiple individual models, but in a sequential order. It learns to reduce predictive error from previous models by modifying the original dataset weights to decrease model biasedness.

g) GridSearchCV: It is a cross-validated method which search through the best parameter values from a given set of grid parameters.

h) RandomSearchCV: It is also a cross-validated method, but instead of searching the parameter it only selects a random parameter and tests a random combination.

i) Artificial Neural Network(ANN: The goal of ANN is to replicate the neural networks similar to human brains so that computers can learn new information in more complex way.

4 Experiments/Results/Discussion

We started experimentation with linear models and followed by tress methods, ensemble methods and neural networks. Our primary metric for model evaluation was accuracy. We got an accuracy of 0.793 in Logistic Regression model. For regularization, we also evaluate mean-squared-error for Ridge, Lasso and ElasticNet. Lasso has mse of 0.162, Ridge has mse of 0.1414 and ElasticNet has mse of 0.1413. ElasticNet has the least mean-squared-error with the value of 0.1413. ElasticNet is more flexible between Ridge and Lasso. Since, lambda is shared penalization parameter while alpha sets the ratio between L1 and L2 regularization. For tree models, the parameters were choosen with the help of GridSearchCV in combination with the above mentioned methods. We choose GridSearch becuase it chooses the best parameter in a given grid. We choose the gini-index to decrease the level of entropy from the dataset. We also did some visualization (like correlation-matrix, heatmaps, pair-plots etc.) and many more insights which are available in the jupyter notebook.

Model Type	Accuracy
Logistic Regression	0.7930
Bagging Model	0.8490
Boosting Model	0.8495
XGBoost	0.8515
KMeans2	0.5405
Decision Tree	0.7785
Tuned Decision Tree using GridSearch	0.8546
Random Forest	0.8515
Tuned Random Forest using GridSearch	0.8564
ANN Train Accuracy	0.7971
ANN Testing Accuracy	0.7929

From the above table, we can see that Tuned RandomForest using GridSearch has the best accuracy score of 0.8564. Because the random forest collects the data of each tree and forecasts the future based on the majority of predictions, rather than relying on a single decision tree. Furthermore, tuning RandomForest using GridSearch just boosts its accuracy little more. we got the lowest accuracy on KMeans-2 clustering because it has trouble clustering data where clusters are of varying sizes and density.

5 Conclusion/Future Work

From the various experiments we have so far, we concluded that a new customer has higher chances of leaving the bank as compared to already existing customer. When a customer has more NumOf-Products they are less likely to exit the bank. In the above table, we can see that Tuned RandomForest using GridSearch has the best accuracy score of 0.8564 outperforming other models. This is because RandomForest collects the data of each tree and forecasts the future based on the majority of predictions, rather than relying on a single decision tree. Furthermore, tuning RandomForest using GridSearch just boosts its accuracy little more. we got the lowest accuracy on KMeans-2 clustering because it has trouble clustering data where clusters are of varying sizes and density. For future, if we have more time, we can further expand our model accuracy using parameter tuning. We would also like to implement the neural networks a bit more effectively by adding more hidden layers and choosing a appropriate activation function to increase the accuracy.

6 Contributions

For this project, we both just sit together on zoom call and discussed which methods/ models would be more appropriate for this dataset. We also brought our experience from previous classes to add more to this project. Data Visualization was helpful to see, interact and better understanding of data (like features). Data Mining was helpful to perform data preprocessing, exploratory data analysis and to extract some useful insights. Most of the machine learning methods that were implemented in the project were introduced in the DataCamp course and assignments which were given in Machine Learning class.

References

The dataset was obtained from "https://www.kaggle.com/datasets/filippoo/deep-learning-az-ann".