Loaning during a Pandemic

Full Data Mining Analysis

Srujana Turaga, Rahul Sai Mallam, Louis Crouch

December 08, 2021

ISM6136 University of South Florida

Contents

Background of problem - Srujana

A bank's capital or net worth determines how well it is doing. It is calculated using a balance sheet.

The balance sheet comprises assets and liabilities for the bank.

Assets are something of value that the bank owns which can be used to produce something such as cash in its vaults and money that the bank holds at the Federal Reserve bank, loans that are made to its customers, bonds, etc. Liabilities are a sum of what is owed. When people deposit money in the bank, it increases its cash flow and credibility with investors however, on the bank balance sheet it is technically considered a liability since the bank must repay the money when the customer wants it.

So how does the bank make money?

Banks charge a higher interest rate on loans than the interest rate they offer in deposits and generate income from the interest rate spread. This is how they primarily make a profit. There is something known as interest rate risk which is the management of the spread between the interest earned through loans and that what is paid in deposits. Deposits are generally short-term and are adjusted to current rates faster than the loan rates. So, if interest rates are increasing then banks can charge a higher rate on both fixed and variable loans. However, if interest rates begin to fall, banks are at a higher risk as their income through interest will decrease.

Motivation for solving the problem - Srujana

The COVID-19 pandemic created a huge unemployment gap across the world and significant shock across various sectors including the Portuguese bank which we have considered in this project. Many of its customers either lost their job or had medical bills to pay due to which they withdrew money from their deposit accounts while others did not complete the monthly loan payment installation. Due to these unforeseen circumstances, the bank's capital dangerously dropped and if continued, would soon lead

them to a loss. The bank's CEO wants to ensure that this does not happen. The bank desires to have a better understanding of these clients as to be able to then gain additional clients to offer loans to. With the bank being able to have a better understanding of their clients, it will allow for a better marketing strategy. That strategy can then address the unique situation of their clientele, increasing the chances of a loan being taken. This would then allow the bank to gain more interest from customers than if it did not market to them.

Solution Methodology and Evaluation Metrics. – Srujana

Data mining is a useful tool that helps identify hidden relationships and trends in your data. It also helps predict the potential results of different data analysis applications. The similar scenario is being used to predict whether a customer would like to take a loan from the bank or not. It is most important to identify the dependent and independent variables depending upon the data, type of problem and the analysis being carried on. In our analysis, the dependent variable is 'Loan'.

So, we went for Two-class classification algorithms which are supervised learning algorithms. We used 80% of the data for training a model and 20% of the data for testing a model. Four different algorithms, Two- Class Decision Forest (which is highly accurate and faster in training times) with varied parameters, Two- Class Neural Network, Two-Class Averaged Perceptron and Two- Class logistic regression (linear and faster in training times) are used to find the most suitable algorithm.

The software we are using is "Microsoft Azure ML (Machine Learning) Studio" which has varied machine learning algorithms to perform, visualize and compare the results.

Here, Accuracy and Precision of these models will be compared to evaluate effectiveness of the model. To be more precise we can take F1 score of each of the model. F1 is an overall measure of a model's accuracy that combines precision and recall. A good F1 score means that we have low false positives and low false negatives.

Description of Dataset – Louis

The dataset that was used for this project was found on Kaggle. The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. This data ordered by date (from Dec 2019 to November 2021).

This dataset consists of material detailing the information for individual clients of a bank. This data consists of key attributes that can affect whether a client would need a loan, based on who has taken a loan. Major contributors to this are: "age", "job", "education", "default status", and their "balance". The month, "pday" and "poutcome" data are critical in understanding the rate at which clients are contacted as regards their loaning history. The campaign field will also keep the bank aware of the frequency of times the bank has "reached out" to the client. The "balance" is also a critical aspect of the bank understanding its clients. How much money does the client have in their balance? This helps make determinations in the algorithm section.

Client Banking Data - Louis

- age (numeric)
- job (categorical)
    - ("admin.","unknown","unemployed","management","housemaid","entrepreneur","student", "blue-collar","self-employed","retired","technician","services")
- marital (categorical)
    - "married","divorced","single"; note: "divorced" means divorced or widowed)
- education (categorical)
    - ("unknown","secondary","primary","tertiary")
- default (1 or 0)
    - has credit in default?
- balance (numeric)
    - Their average yearly balance, in dollars.
- housing (1 or 0)
    - Is there a housing loan?
- contact: contact communication type (categorical: "unknown","telephone","cellular")
- day:
    - last contact day of the month. (numeric)
- month (categorical)

    Last contact month of year "Jan", "feb", "mar", …, "nov", "dec")
- duration (numeric)
    - Last contact duration in seconds.
- Campaign (numeric, includes last contact)
    - The number of contacts made during this campaign and for this client.
- pdays (numeric)
    - The number of days that passed by after the client was last contacted from a previous campaign.

- previous (numeric)
    - The number of contacts performed before this campaign and for this client.
- poutcome (categorical)
    - The outcome of the previous marketing campaign ("unknown","other","failure","success")
- Output variable (desired target)
- Term Deposit (1 or 0)
    - Has the client subscribed a term deposit?
- Loan (1 or 0)

Comparison of Algorithms - Rahul

Classification of Individuals

Figure 1 below, shows a snippet of workflow defined in AzureML studio. We tried two different models to Left side flow we used Two-Class Decision Forest and Two-Class Neural Network. To the right we used a Two-Class Average Perceptron and Two-Class Logistic Regression.

The dataset we are processing is clean and understood. Before processing, the data need to load into the AzureML studio and need to be split into two different sets.

1. Training Set (80% of total data)

2. Testing Set (20% Remaining data)

The data was sent to the four algorithms:

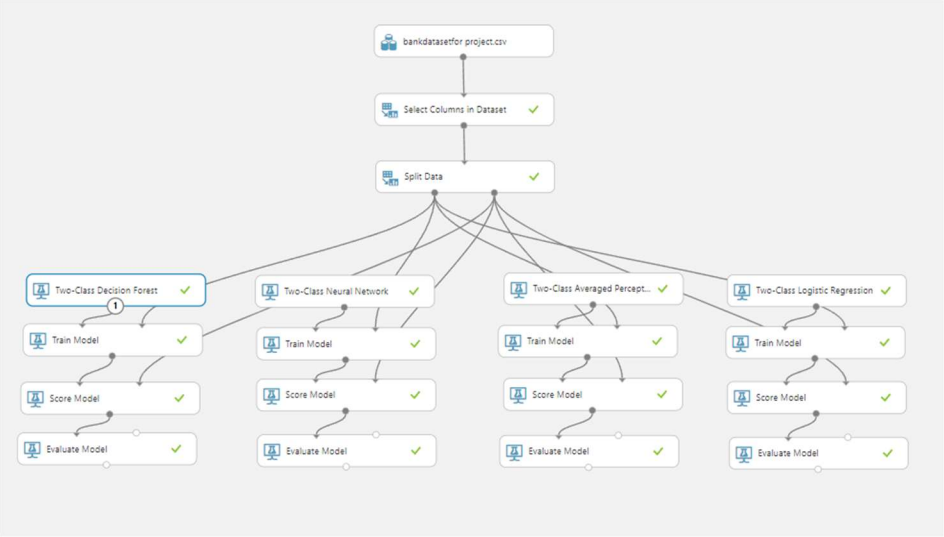
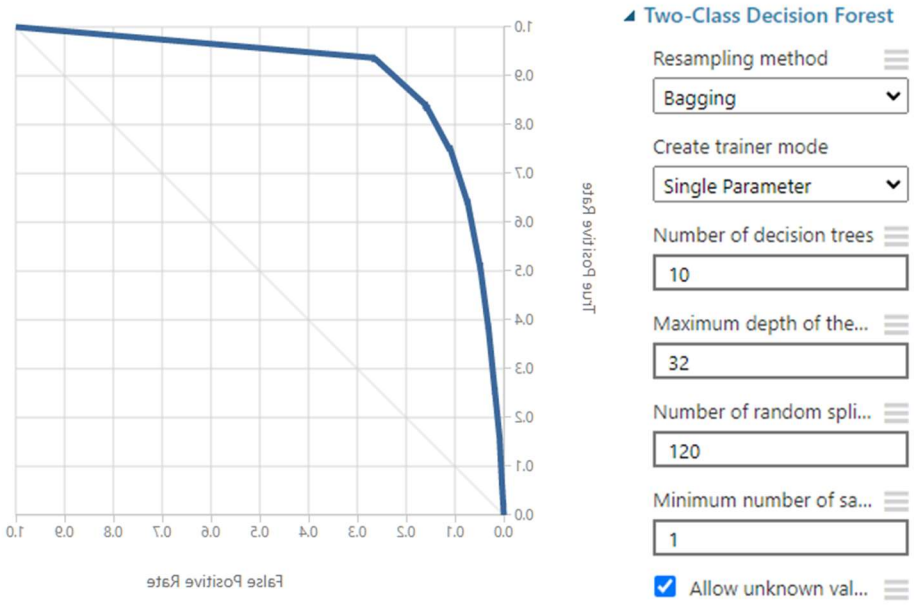


Figure-1 - AzureML Workflow applying Two-Class Decision Forest, Two-Class Neural Network, Two-Class Average Perceptron, and Two-Class Logistic Regression.

Two-Class Decision Forest

The term Decision Forest regression is ensemble of decision trees. The machine learning model based on random decision forests algorithm. The main usage of Decision Forest regression is to predict the target variable which has two values.

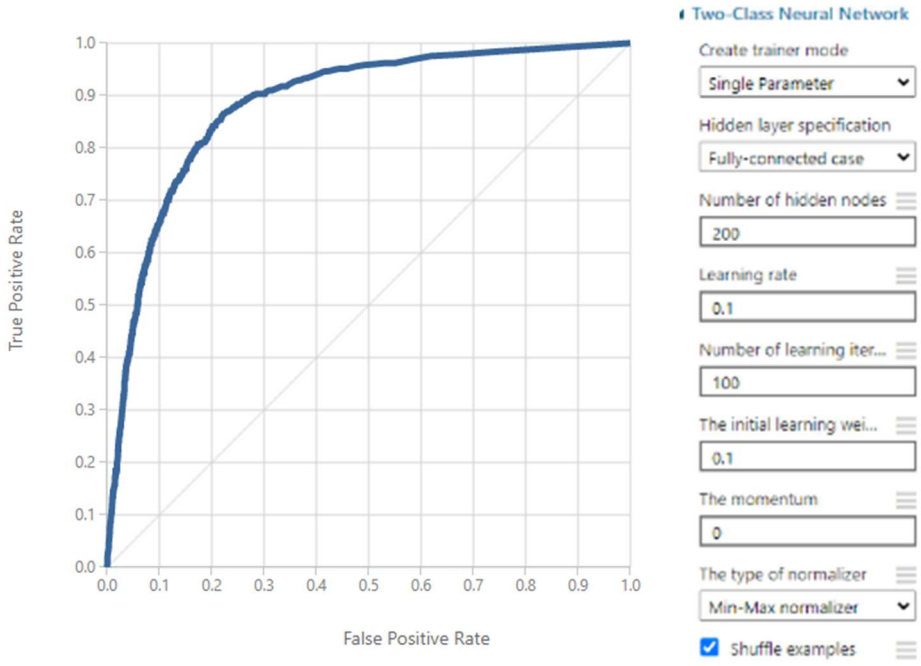


True Positive	False Negative	Accuracy	Precision
408	644	0.901	0.615
False Positive	**True Negative**	**Recall**	**F1 Score**
255	7735	0.388	0.476

Here, we are using 10 decision trees and maximum depth of the decision tree is 32, Remaining all the values are set to default.

Two-Class Neural Network

Neural network is a complex machine learning used to identify patterns. It contains input, hidden and output layers. Neural networks are trained iteratively using optimization techniques called gradient descent. Back propagation is an iterative algorithm that modify the weights of a neural network little bit at a time.



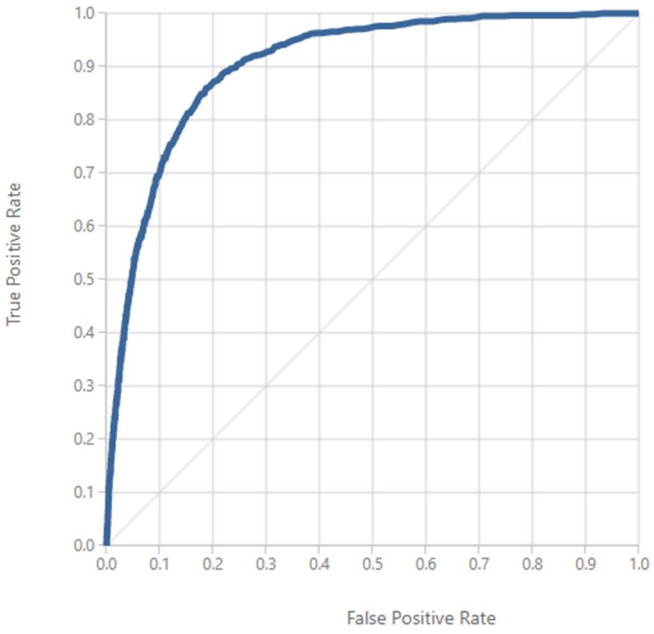
True Positive	False Negative	Accuracy	Precision
376	676	0.895	0.581

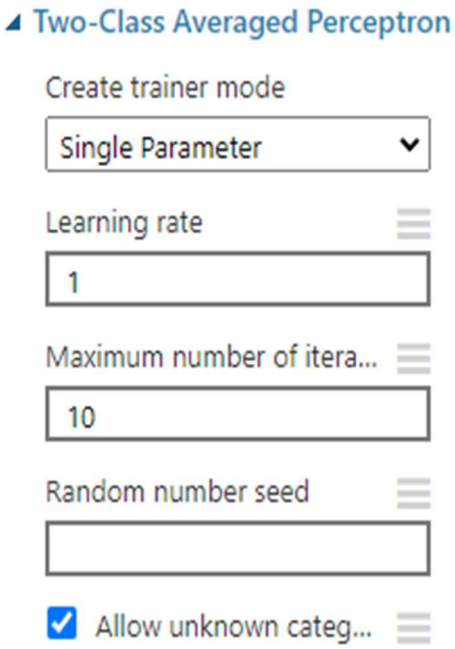
False Positive	True Negative	Recall	F1 Score
271	7719	0.357	0.443

In this algorithm, we are using 200 hidden nodes.

Two-Class Average Perceptron

The averaged perceptron approach is an early and basic version of a neural network. In this technique, inputs are categorized into multiple possible outputs based on a linear function, and then coupled with a set of weights that are obtained from the feature vector.

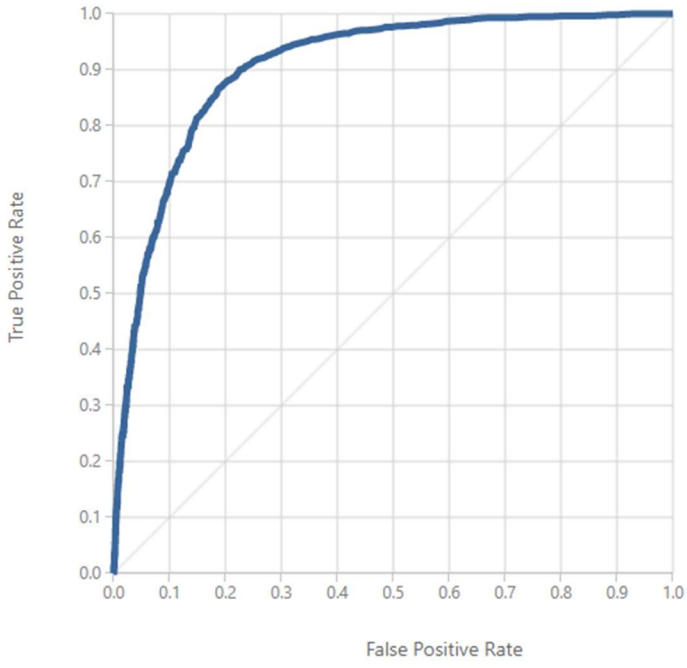


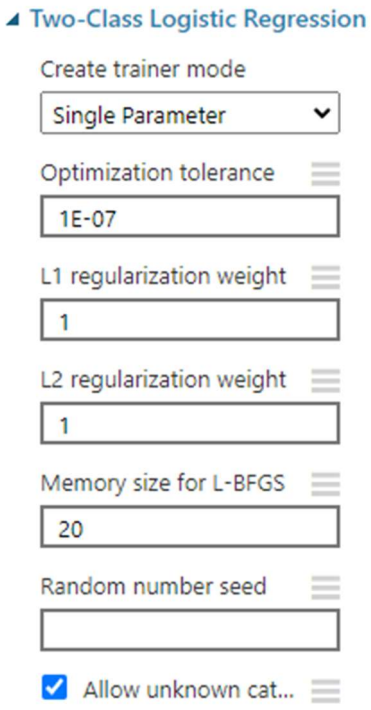


True Positive	False Negative	Accuracy	Precision
235	817	0.897	0.675
False Positive	**True Negative**	**Recall**	**F1 Score**
113	7877	0.223	0.336

Two-Class Logistic Regression

It is a predictive analysis algorithm which is used between dependent variable and one or more independent variables to predict the probability of target variable. Mostly it is used for classification problems.





True Positive	False Negative	Accuracy	Precision
317	735	0.899	0.642
False Positive	**True Negative**	**Recall**	**F1 Score**
177	7813	0.301	0.410

Summary sheet showing the results of all experiments – Rahul

Now that we have run different algorithms, we have different values generated for different metrics such as precision, recall etc.

Precision: Is as the number of true positives divided by number of true positives and false positives. It is also called as positive predictive value. The precision value is affected from false positive from confusion matrix.

$$P = TP/(TP+FP)$$

Recall: Is the number of true positives divided by number of true positives and false negatives. It is also known as true positive rate. The recall value is affected from false negatives

$$Recall = TP/(TP+FN)$$

Accuracy: Is the sum of true positives and true negatives divided by true positives, true negatives, false positives, and false negatives.

$$Accuracy=(TP+TN)/(TP+TN+FP+FN)$$

F1 Score: It shows the balance between precision and recall.

$$F1\ Score = 2*((Precision*recall)/(precision+recall))$$

All the results of the experiment are listed below in a tabular format.

ALGORITHM	ACCURACY	PRECISION	RECALL	F1 score
Decision Forest	0.901	0.615	0.388	0.476
Neural Network	0.895	0.357	0.581	0.443
Average Perceptron	0.897	0.675	0.223	0.336
Logistic Regression	0.899	0.642	0.301	0.410

From the above results, based on the Accuracy and F1-score calculated we can conclude that the **Two-class Decision Forest** with 10 decision trees and maximum depth 32 is the best model we can use to classify customers into their categories. This model has the highest accuracy of 90.1% and F1 score of 0.476

From the confusion matrix, we can say that **408** customers (I.e., True positives) are highly likely to take the loan and **7735** customers are highly unlikely to take the loan.

Precision rate for this model is **61.5%** and Recall rate is **38.8%** which are consistent.

Now that we have a model to predict the customers who take the loan, now we must make recommendations for the bank to increase its Net worth.

Conclusions and Recommendations – Louis

We conclude that based on the data collected, we will be relying on the Decision Forest model, with an accuracy of 90.1%, to analyze the data that we have collected. With this the bank more easily understands which clients that they will need to market loans to. Having a better understanding of their clients can allow for better marketing campaigns that can address demographics of their clientele, increasing the chances of a loan being taken. The marketing should be directed toward specific clients. This can increase the likelihood that the client will accept the loan offer that is made by the bank. This can also be more probable for certain income groups during the COVID-19 pandemic. For those clients who are less likely to take a loan, we recommend conducting a survey to gather more information which could be used in the future. This information can help the bank have a better understanding of how to market to these different groups.

We recommend that the bank adopt direct marketing to its individual clients. This can be a phone call, text, or email. We also recommend that if there is a risk based on the income level of the client, or past credit history, a margin can be added on the interest that they take, and the bank can keep a separate cash reserve specially dedicated to loan defaults. This secures the bank so that it can cover a loss if the client is unable to pay back their debt to the bank.

Overall, the results that the decision tree produces will allow us to have a good understanding of a customer's behavior, which in turn will allow the bank to make better decisions that will ensure its capital adequacy, even during tough economic times. In this case, a global pandemic.

(This paper reviews what the bank would be doing at the beginning of the COVID-19 Pandemic. This though is negated by the Federal CARES Act which was signed into law on March 27, 2020. This proposal stands as well in the event of the CARES Act not being signed into law.)

References

Jha, S. (2018, October 23). Bank Marketing. Kaggle.com. https://www.kaggle.com/sonujha090/bank-marketing

Moro, S., Raul M S Laureano, & Cortez, P. (2011). USING DATA MINING FOR BANK DIRECT MARKETING: AN APPLICATION OF THE CRISP-DM METHODOLOGY. http://repositorium.sdum.uminho.pt/bitstream/1822/14838/1/MoroCortezLaureano_DMApproach4DirectMKT.pdf