CAPSTONE PROJECT, Loan Default Prediction

Using machine-learning model techniques to predict clients who default on loan

By: Erik Guevara

22 January 2022

EXECUTIVE SUMMARY

Introduction: A major proportion of retail bank profit comes from interests in the form of home loans. Banks are most fearful of nonpayers, as bad loans usually eat up a major portion of their profits. Therefore, it is important for banks to be judicious while approving loans for their customer base. The objective is to build a classification model to predict clients who are likely to default on their loan and give recommendations to the bank on the important features to consider while approving a loan.

Data Collection: Data used for this analysis comes from The Home Equity dataset (HMEQ). Data was downloaded with Python programming language.

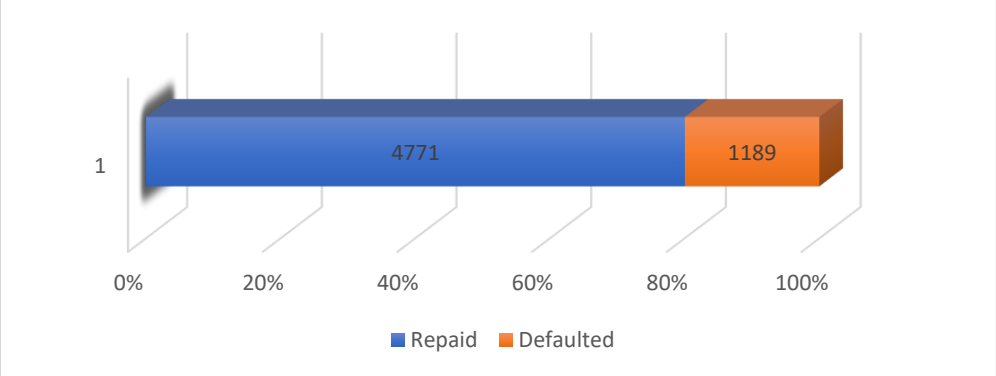
Exploratory Analysis: Exploratory analysis was performed in order to identify the quality of data (such as identifying and treating extreme values and missing values) and to determine relationships between all the variables.

Results: The dataset used in this analysis contains 5,960 observations and 13 variables:

BAD: 1 = Client defaulted on loan, 0 = loan repaid. **LOAN:** Amount of loan approved. **MORTDUE:** Amount due on the existing mortgage. **VALUE:** Current value of the property. **REASON:** Reason for the loan request. (HomeImp = home improvement, DebtCon= debt consolidation which means taking out a new loan to pay off other liabilities and consumer debts), **JOB:** The type of job that loan applicant has such as manager, self, etc., **YOJ:** Years at present job., **DEROG:** Number of major derogatory reports. **DELINQ:** Number of delinquent credit lines. **CLAGE:** Age of the oldest credit line in months. **NINQ:** Number of recent credit inquiries. **CLNO:** Number of existing credit lines. **DEBTINC:** Debt-to-income ratio.

The target (BAD) is a binary variable that indicates whether an applicant has ultimately defaulted or has been severely delinquent. This adverse outcome occurred in 1,189 cases (20%). 12 input variables were registered for each applicant.

The intended goal is to identify those applicants who have high risk of defaulting or being delinquent, by identifying those applicants we can stop from providing loans in the first place and therefore minimize losing resources.



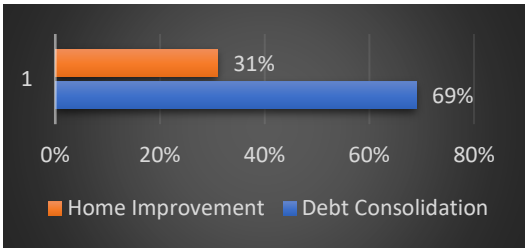
We use different approaches to analyze the variables, we analyze them individually and in pairs to assess the relationship between them and understanding the influence they have in those customers who defaulted on loan.

Why is this problem important to solve?

Defaulted or severely delinquent applicants translates into lost resources for the loan entity, it's a priority to minimize those loses so the company can keep generating income and grow.

For those who cannot pay back the loan, what variables they have in common? What patterns we can observe that allow us to put that group together so we can manage better the loan approval process?

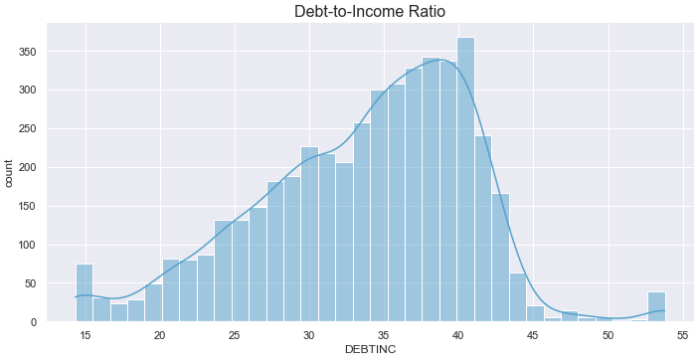
With Exploratory Data Analysis and Visualization, we are getting insights, we are extracting valuable information that might not be easily interpreted by reviewing the raw data. With Machine learning we can build a model that allows to predict if a potential applicant will pay back the loan or default it. Using EDA and machine leaning we have the tools for better decision making that would make the company more efficient managing the resources.

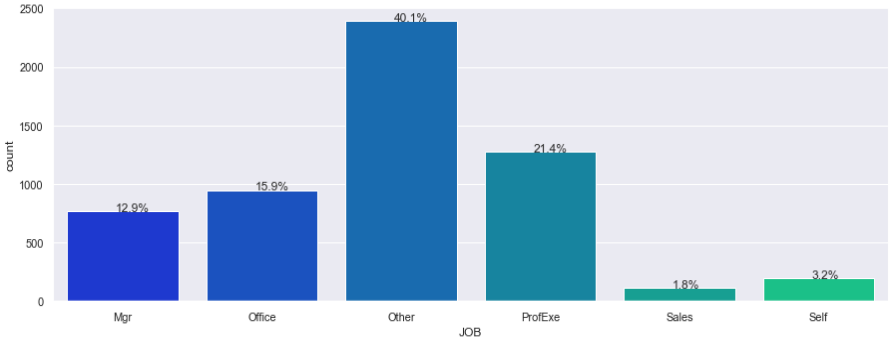


The variable REASON has only two options, the applicant needs the loan either because they want to perform a home improvement or because they need to consolidate the debt, around 2/3 parts of the applicants request a loan so they can consolidate the debt.

In terms of who defaulted and who repaid, there is no significance difference, in both cases the percentage is similar, therefore we can say that the variable REASON is not a factor to determine who will or will not repay the loan.

In the debt-to-income ratio variable we observe a left skewed distribution plot with high central tendency values. This is a very important variable to consider for the loan institutions since it measures what is the relationship between the income and the debt, therefore higher this number harder is for the customer to pay back, lower this value more chances loan companies accept the applicants, in our dataset the average is 33.7%, however, the accepted industry standard for debt-to-income ratio is below 20%.

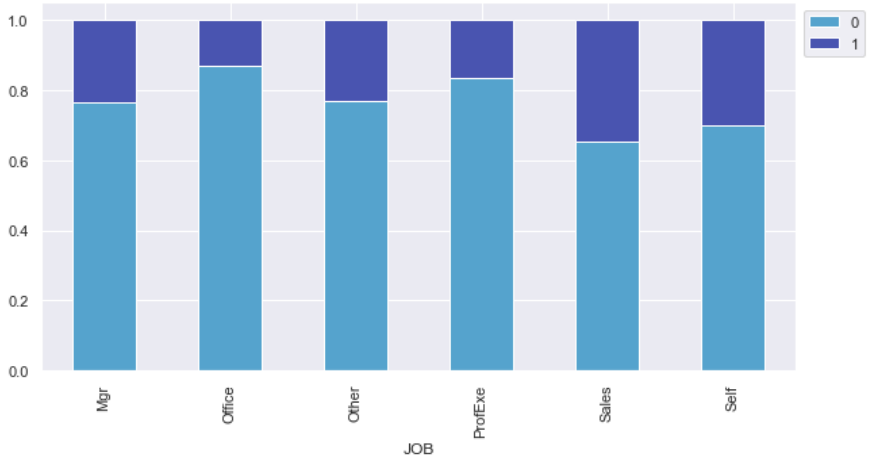


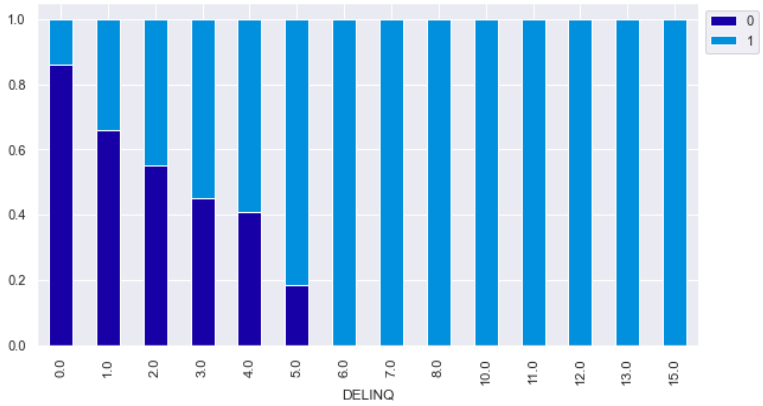


The JOB variable comes with 6 different options, however, one of those options is a very general category called “Other”, this category has over 40% of the observations, it’s unknown how many other jobs are included within this category. We can observe that the “ProExe” has over 20% of observations, this is the second biggest job category; the one with less observations is “sales” with less than 2% of the observations.

What is the relationship between the type of job and defaulted customers?

In the following chart we can observe that some job types have more chances to default than others, the light blue section represents clients who repaid and the purple the clients who defaulted the loan, we can see that those applicants who belong to an “Office” category have a better repay ratio, that is around 86% versus around 65% of sales employees, this shows that the applicant profession plays an important role. “Sales” and “self” categories seems to have the worse repay ratio but those categories also have the least clients in the dataset.

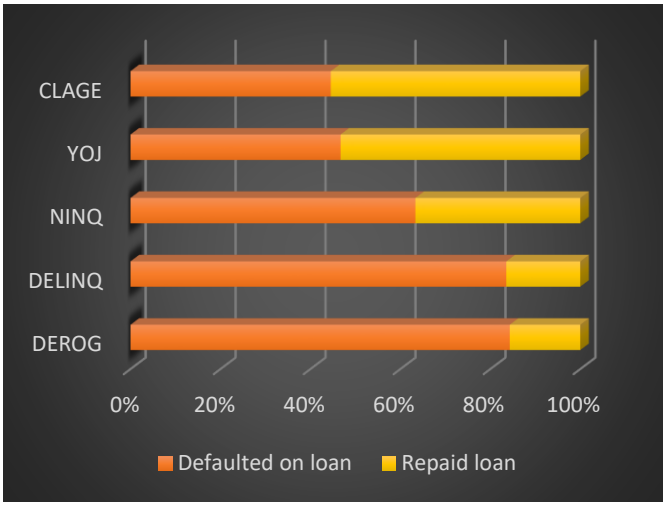




We have analyzed the existing relationship between the client's delinquent credit lines and the defaulted clients and we found the following:

In light blue we can see the defaulted clients and in dark blue we can see clients who repaid the loan. As clients get more delinquent credit lines, they have more chances to default on loan. On the other hand, clients with less delinquent credit lines the chances of returning the loan are considerably higher, therefore this is a very important factor to consider. Those with 0 delinquent credit lines repay the loan, those with 5 delinquent credit lines the ratio is the opposite.

We analyze all the variables against the "BAD" variable, however, we put together specific 5 variables that we consider explain better defaulted customers, we averaged each one and we split them into defaulted on loan (orange) and repaid loan (yellow) and we get the following findings:



CLAGE. The average age of the oldest credit line is lower in the defaulted group, that means that defaulted clients cannot keep credit lines open for longer periods, we can assume they cannot pay on time and their credit lines were close and they need to open new accounts.

YOJ. The average years at present job is lower in the defaulted group.

NINQ. Defaulted customers have in average a higher number of recent credit inquiries; this possibly indicates that the defaulted customer was trying to get new credit lines somewhere else, presumably they cannot open new credit lines because a negative record.

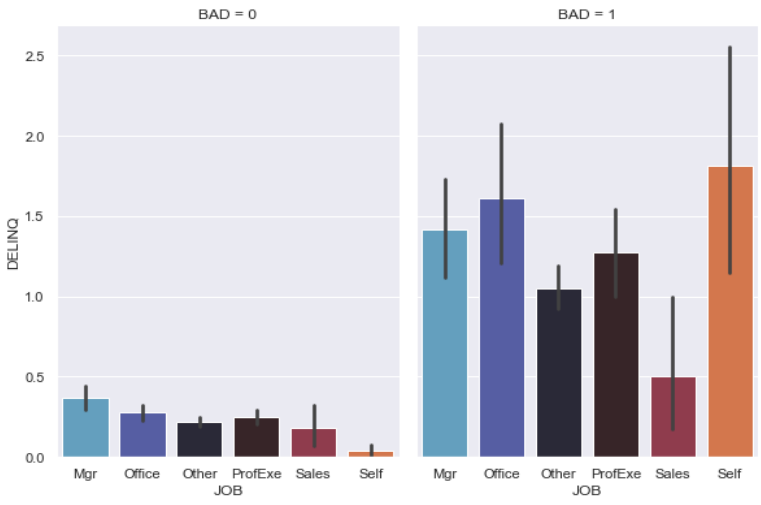
DELINQ. We can observe the strong relationship between defaulted customers and the average number of delinquent credit lines. An account become delinquent when the customer cannot pay on time. As we can see, the defaulted customers have noticeably more delinquent credit lines in average than a customer who paid the loan.

DEROG. The average on derogatory reports is considerable higher on the defaulted group. Serious delinquency or late payments are reported and on average defaulted customers have a prominently more of those reports.

Here we observe the relationship between 3 different variables: defaulted clients, number of delinquent credit lines and the job the client does. we can observe how the type a job acts differently on each case

For example, defaulted self-employees have a considerably more delinquent credit lines compared with those self-employees who repaid the loan, that's a red flag we can put in consideration.

We can observe how office workers tend to have high delinquency on the defaulted group and just average on the repaid group.

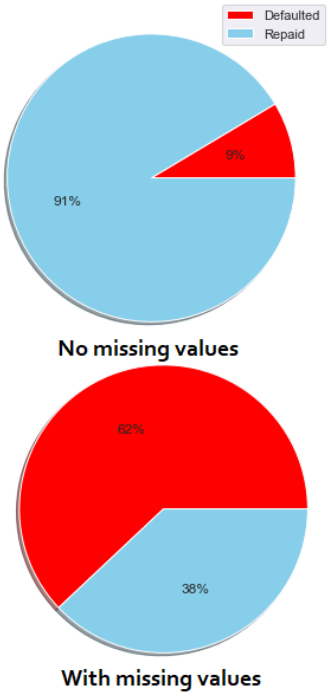


So far, we have discovered that out of the 12 variables, 5 of them play an important role to determine who might default on loan which are CLAGE, YOJ, NINQ, DELINQ and DEROG, being extremely important the last two.

Understanding observations on missing values

The dataset comes with an important number of missing values, every single variable except for BAD and LOAN has a certain degree of missing information, this could be information not entered by the customer or not entered by the person who get the information. The most missing values were found in the Debt-to-income ratio variable, with over 20% of missing information, this is 1 in 5 observations lack of this important data.

We separated the dataset in 2 parts, 1 subset that includes all the Debt-to-income ratio variable with no missing values, and another subset with the same variable but with missing values and when we compare both subsets against defaulted clients, we get the astounding result on the right.



The one at the top represents the customers who Debt-to-income variable has been filled out (no missing values) and at the bottom we have the customers with Debt-to-income variable empty (missing values); the defaulted customers in red and the repaid customer in light blue.

We can observe a very important difference; on the top plot we can see that the majority of customers who completed the DEBTINC data repaid the loan, ~91%, on the other hand, at the bottom, we can see that more than half of customers who didn't complete the DEBTINC data unsuccessful pay back the loan.

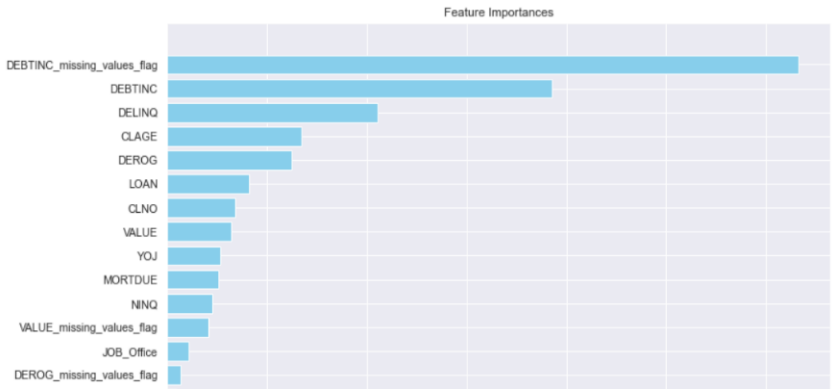
We can assume that most of the customers who have the tendency of not paying back the loan preferred not to disclose the Debt-to-income ratio information. Possibly they have such a high Debt-to-income ratio that they are afraid of being rejected and decide not to include this important information.

Missing values on DEBTINC plays an important factor to determine defaulted customers on loan.

After finding this result, we realize how important and impactful could mean having missing values, therefore we added those missing values to the equation and for each column we created a binary flag. A flag variable will be added to the dataset so we have a better data to be trained by our machine learning models.

REFINED INSIGHTS

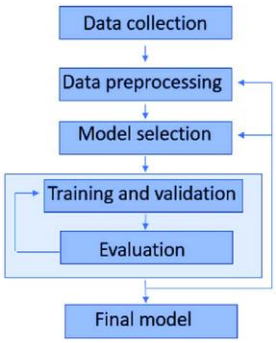
After Exploratory Data Analysis we could identify those variables that helped us to understand those customers who default and those who repaid, after a deeper analysis and using machine learning models we could redefine those findings confirming some variables that we found initially with EDA but also discovering other variables that without machine learning tools would have been tremendously difficult.



Out of the 11 variables in the dataset, 4 of them play an important role to understand much better the relationship between them and defaulted customers.

- DEBTINC (Debt-to-income ratio)
- DELINQ (Number of delinquent credit lines)
- DEROG (Number of major derogatory reports)
- CLAGE (The type of job that the loan applicant has)

What is the potential solution design?



After data collection and data processing the next step is to select the correct machine learning model**,** Since the problem to solve is a classification and it it's presented in binary form, we will focus on those machine learning models that deals better with categorical data such as regression models and random forest. The objective will be predicting applicants who are prone to default.

After model selection, we have trained and validate the information for evaluation, most of the models were trained with a variety of hyperparameters and different variations of the same dataset with different missing values and outliers' treatments, the hyperparameters were training with different weight ratios, split ratios, techniques for imbalanced data, different criterion values, different maximum depth and minimum samples splits values; all of them providing different results in the train and test data.

After evaluation we have returned to previous stages of the process like data processing and model hyperparameters adjustments, the focus is to predict accurately defaulted customers. We iterated the process several times until we reached the best results in each machine learning model and at the end we compared the results to focus in the final model.

COMPARISON OF TECHNIQUES AND THEIR PERFORMANCES

Since the problem to solve is a classification, we have focused on those machine learning models that deals better with categorical data, we trained models and compare the following machine learning techniques:

Logistic Regression | Decision Tree with weight | Decision Tree with hyperparameters | Random Forest | Random Forest with weight | Random Forest with hyperparameters

Model Evaluation Criterion

How to reduce the losses? We use the confusion matrix to identify the false negatives which are our target. The value to observe is the recall since it represents how accurate the model is at predicting those customers who will default. The organization would want Recall for class 1 (defaulted customers) to be maximized, greater the Recall score higher are the chances of minimizing False Negatives.

Results on each machine learning model:

Logistic regression, recall 59%

- The accuracy in both the test and training show over 88%, however, the recall on class 1 below 60% this will produce negative results at identifying defaulted customers.
- According to the coefficients the Value of the property on missing values is the most important factor, followed by DEBTINC with missing values and CLNO with missing values.

Decision Tree with weight, recall 63%

- Class weights tries to balance the desired results in one particular class, on this case since the class 1 is our priority we set this parameter with 80% of weight.
- The Decision Tree works perfectly on the training data but not as accurate on the test data as the recall is 0.63 for class 1 as compared to 1.00 for the training dataset.

Decision Tree with hyperparameters, recall 86%

- With the Hyperparameter Tuning the Decision Tree shows worse performance on the training data compared with the Decision Tree with weight parameters, the recall is 0.89 for class 1 as compared to 1.00 using the weight model.
- The Hyperparameter Tuning Decision Tree performance in the test data is slightly worse than the training data, however, the recall is 0.86 for class 1 as compared to 0.89 obtained in the training dataset, however, this recall is higher than any previous model, therefore this model with the Hyperparameter Tuning shows the best result so far**.**
- This hyperparameter has maximized the recall.

Random Forest, recall 71%

- Random Forest is an algorithm where the base models are Decision Trees. The results from all the decision trees are combined together and the final prediction is made using voting or averaging.
- The Random Forest model works perfectly on the training data but not as accurate on the test data as the recall is 0.71 for class 1 as compared to 1.00 for the training dataset.
- The recall on the test data is inferior compared to the decision tree with hyperparameters 71% vs 86%.

Random Forest with weight, recall 70%

- The Random Forest with class weight is giving 100% results on the training dataset, 100% accuracy and recall.
- The recall on the test data is inferior compared to the random forest with no class weight.

Random Forest with hyperparameters, recall 82%

- The recall value is the higher number observed on a Random Forest, however, the value is not as high the hyperparameter decision tree model.
- The Random Forest shows a slightly better overall performance, but the defaulted prediction factor is better captured with the hyperparameter decision tree.

In terms of overall performance, the random forest with weight parameters shows the best results predicting with high grade of accuracy both classes, default and repaid customer, however, the recall value is not as good compared with other models.

The logistic regression model shows overall the worse performance, nevertheless every single model provided valuable information difficult to obtain just by performing Exploratory Data Analysis.

Comparing Model Performances:

Model	Train Accuracy	Test Accuracy	Train Recall	Test Recall	Train Precision	Test Precision
Logistic Regression	88%	89%	60%	59%	78%	81%
Decision Tree Weight	100%	88%	100%	63%	100%	73%
Decision Tree Tuned	84%	84%	88%	86%	56%	56%
Random Forest	100%	91%	100%	71%	100%	86%
Random Forest Weight	100%	91%	100%	70%	100%	87%
Random Forest Tuned	88%	88%	84%	82%	66%	68%

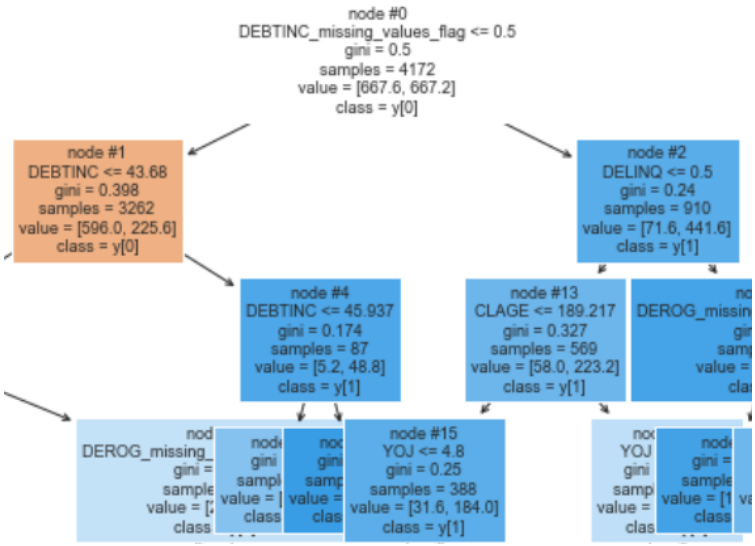
PROBLEM AND SOLUTION SUMMARY

Detecting and predicting what customer will default the loan is the problem being solved, the exploratory data analysis helped to identify those important variables, but machine learning models find other relationships and this combined information is valuable for the decision making.

I propose the Decision Tree with hyperparameters. This model is solving the problem of detecting defaulted customers. This model shows an 86% of recall follow by 82% using the Random Forest with hyperparameters. The advantage of this model is that it will predict mode accurately those customers who might default, the disadvantage is it won't predict well the customer who will repaid as compared with a Random Forest but since our priority is predicting the defaulted customers, we have to sacrifice some precision on the repaid customer class.

The second advantage of the Decision trees is its interpretability, because of the nature of the decision tree graph is relatively easy to follow the nodes and determine the likelihood of being a default or repaid customer based on the flow of those nodes as show below:

- We can see that the first node is based on debt-to-income ratio with missing values, this is the initial factor to determine who might default the loan, if the customer has a missing value on this variable, then it moves to the right of the graph.
- Based on the color coded we have a visual aid that shows that the left side has an orange leaf, showing customers who repaid the loan, on the other side we can see blue leaves indicating customers who defaulted on loans.
- In the second node for those who have no missing values on DEBTINC the next condition is if the DEBTINC is less than 43.68, if this is true once again the tree moves to the blue area and the next validation attribute is again DEBTINC.
- As the tree moves below, we encounter more conditional variables that determine if the customer will default.



RECOMMENDATIONS FOR IMPLEMENTATION

Recommendations for the current data base:	Recommendations for collecting information:
Using Machine learning models: 1. Deploy the suggested machine learning model to automate the prediction of defaulted customers. 2. Re-train the machine learning models with periodicity so that the company has the most accurate model. Using filtering data: When getting the information from the customer we can use the 4 variables described on the future importance, based on those values the applicant could be accepted or rejected based on how high or low are those values. It is suggested to filter the information and determine who is having high numbers that are interpreted as red flags, for instance, the Derogatory report number, the Number of Delinquent credit lines, the Age of the oldest credit line. The finance company should manage the credit risk by filtering out those customers with high Debt-to-Income ratios as they might run into trouble repaying their loan in case of financial hardship. If the company still decides to provide with a loan, they can Increase the interest rate for those customers with more chances of being defaulted.	1. When getting information from applicants, all the fields must be mandatory instead of optional, that way we make sure customers skip or miss important information, particularly the debt-to-income ratio, as we saw previously, a big part of defaulted clients didn't add this information. 2. Instead of having the debt-to-income ratio, it's suggested to create 2 variables, one for income, and another one for debt, with this suggestion we have 2 valuable information which we can use to create our own debt-to-income variable. 3. Diversify the "JOB" variable, instead of "Others" it is suggested to add a wide range of occupations, that way in the future we can understand much better the relationship between jobs and defaulted clients and integrate that into future machine learning models. 4. Customers might enter fake information to increase the chances of being accepted, is suggested to validate all the information. By applying those recommendations, the company will get not only quality in the information, but also accurate data that would lead to build a more precise machine learning models that will help even better to predict defaulted customers and understand much better all the variables.

- By applying these solutions, the expected benefit is the company will reduce greatly defaulted customers and this will allow the company to minimize losses and maximize profits.
- Other benefits include automatize processes, reduce manual analysis, decrease the time to complete the loan process, process more customers in less time.
- The risk with the automated machine learning model is that there will be a small percentage of applicants who will be predicted as defaulters when in reality they will repaid, as more applicants are entered into the system the machine learning model must be retrained, the challenge is to keep the model working at the optimal level so that it delivers the desired outcome.
- The analysis has shown it is feasible to develop a prediction model that makes possible to predict defaulted on loan customers. Acceptable performance in the range of 84% correct classification rate, levels of cross-validation error which seem to indicate the models will be robust to future data sample from similar datasets.
- It's important that the company should invest in resources to keep acquiring quality data from the customers and train the machine learning models constantly to maximize the accuracy.
- Adding and removing other variables often changed the model results, suggesting that there are other interactions in the model that have not been fully explored by this analysis.
- Collecting quality data and further investigation will be necessary in order to explore more fully what factors have predictive effects on defaulted customers.
- It is possible that additional factors not represented in the dataset (individual income, household income, number of children, etc) should be taken into account.